buildings

MDPI

Article

An Immersive Audiovisual Simulation Framework Based on 360° Video and Spatial Audio for Built Environment Perception Evaluation

Yaqi Pan [1,†], Huangyan Zheng [2,†], Jian Xu [3], Peiyuan Yang [2], Peijie Liu [2] and Shuo Wang [2,*]

[1] School of Architecture, University of Sydney, Camperdown, NSW 2006, Australia; ypan0312@uni.sydney.edu.au
[2] School of Architecture, South China University of Technology, Guangzhou 510641, China; hyan40517@gmail.com (H.Z.); yangpeiyuan1009@163.com (P.Y.); pjliu@scut.edu.cn (P.L.)
[3] Transportation and Municipal Engineering Institute, Zhejiang Key Laboratory of Intelligent Control of Transit Infrastructure Risk, Power China Huadong Engineering Corporation Limited, Hangzhou 311122, China; xu_j29@hdec.com
* Correspondence: shwang@scut.edu.cn
† These authors contributed equally to this work.

Abstract

Conventional laboratory-based soundscape evaluations often rely on static images and audio playback, which may provide limited visual dynamics, spatial auditory information, and interactive experience. This study developed an interactive audiovisual simulation system that integrates 360° video with spatial audio for controlled soundscape evaluation. Visual stimuli were presented on a monitor, with viewing direction controlled using a mouse in the 360° video condition. A laboratory experiment compared four scenarios combining two visual presentation modes—static photographs and 360° video—with the presence or absence of background sound recorded on site. Spatial audio was used in all four scenarios. Thirty-eight participants were recruited, and data from 36 participants with normal hearing and normal or corrected-to-normal vision were included in the final analysis. Participants evaluated visual perception, auditory perception, spatial perception, perceived realism, and overall perception across four audiovisual scenarios. Wilcoxon signed-rank tests showed that without background sound, the 360° video condition received significantly higher ratings for visual perception, perceived realism, and overall perception. With background sound, significantly higher ratings were observed for visual perception, auditory perception, spatial perception, perceived realism, and overall perception. For overall perception, the effect sizes were r = 0.379 without background sound and r = 0.668 with background sound. These findings indicate that combining panoramic visual presentation with dynamically rendered spatial audio can provide a more comprehensive perceptual representation of environmental scenes than conventional photograph-based methods. The proposed system offers a controllable and efficient approach for multisensory soundscape and built-environment evaluation. Further comparisons with in situ evaluations across different populations and settings are needed to assess its broader applicability.

Keywords: audio–visual interaction; soundscape; 360° video; spatial audio; virtual reality

Check for updates

Academic Editor: Antonio Caggiano

Received: 23 August 2026
Revised: 20 September 2026
Accepted: 25 September 2026
Published: 1 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

With the acceleration of urbanization, the quality of human settlements has received increasing attention from researchers and urban planners. Environmental experience is

essentially a process of multisensory integration, in which vision and hearing are considered two of the most important perceptual channels. Early evidence for this view came from landscape-value studies showing that congruent sound can systematically shift the perceived value of an otherwise identical visual scene [1], while subsequent multisensory assessment work quantified the specific contribution of acoustic perception to overall landscape evaluation [2]. More recent conceptual frameworks have consolidated such findings into descriptor systems for predictive soundscape modelling [3] and into design-oriented accounts of a soundscape's role within the built environment [4]. Traditional environmental assessments often examine visual landscapes and acoustic environments separately, thereby overlooking the complex interactions between them at the perceptual and cognitive levels. In urban open spaces, landscape elements and soundscape characteristics rarely function independently. Instead, they interact to influence people's aesthetic experience, comfort, environmental satisfaction, and behavioral responses [5–7].

Soundscape theory further emphasizes that the acoustic environment should be understood from a human-centered perceptual perspective rather than evaluated solely through objective acoustic indicators such as sound pressure level [8]. Previous studies have shown that natural sounds, birdsong, and music can enhance perceived environmental restorativeness, whereas negative sound sources such as traffic noise may reduce environmental satisfaction [9,10]. Visual information can also influence auditory perception. For example, visible street greenery can mitigate the negative effects of traffic noise [11], while improvements in visual environmental quality can increase perceived soundscape pleasantness [12,13].

However, traditional soundscape research has mainly relied on either in situ evaluation or laboratory simulation using static images combined with audio recordings. In situ evaluation provides a high degree of environmental authenticity but offers limited control over external variables. Laboratory simulation, in contrast, provides stronger experimental control and repeatability but may fail to reproduce the visual dynamics and spatial acoustic characteristics of real environments [14,15]. Therefore, there is a need for an immersive and controllable experimental platform that can simultaneously provide realistic visual representation, spatial audio reproduction, and interactive audiovisual presentation.

1.1. Audiovisual Interaction and Environmental Perception

Previous studies have demonstrated significant interactions between visual and auditory information, which jointly shape perceptions of environmental realism, sense of place, and spatial presence [16,17]. In virtual environments, audiovisual consistency is considered an important factor in immersive experience, whereas mismatched visual and auditory cues can weaken the credibility and immersion of a simulated scene [18].

Natural visual landscapes and natural sound sources, such as birdsong and flowing water, can produce synergistic effects that enhance environmental restorativeness. In contrast, inconsistent audiovisual cues may reduce perceived realism and immersion [19,20]. Potter et al. showed that both visual rendering and spatial audio rendering make important contributions to immersive experiences in virtual reality environments [21]. In particular, spatial audio can strengthen users' understanding of scene structure and directional information, thereby enhancing spatial presence and environmental realism.

Studies of urban parks have also shown that visual elements, including vegetation, water bodies, and landscape openness, interact with sound-source type, rhythm, and pleasantness to influence soundscape experience [2,3]. These findings indicate that visual and auditory information are not processed independently but interact to shape environmental perception. Accordingly, integrated audiovisual evaluation methods are needed to represent real-world environmental experiences more accurately [22]. Although previous

research has established the importance of audiovisual interaction, relatively limited attention has been paid to how different visual presentation formats and spatially responsive audio jointly influence perceptual evaluation in controlled soundscape experiments.

1.2. From Parametric Metrics to Lived Experience: In Situ and Laboratory-Based Soundscape Evaluation

Traditional soundscape research has predominantly relied on in situ evaluation. However, field studies are affected by environmental variables such as noise fluctuations, weather conditions, pedestrian activity, and crowd density, which are difficult to control. Consequently, an increasing number of studies have adopted laboratory simulation methods and compared the resulting evaluations with in situ findings.

Hermida Cadena et al. compared in situ and laboratory-based soundscape evaluation methods in three urban parks and examined the causes of differences between the two approaches [15]. Lian et al. compared field-based soundscape evaluations with laboratory simulations and investigated the similarities and differences between the two methods, including the appropriate sound pressure level for laboratory reproduction [14].

These studies indicate that laboratory environments offer advantages in experimental repeatability and variable control while also highlighting the importance of maintaining sufficient realism and immersion in audiovisual materials. Georgiou et al. further emphasized that credible soundscape reproduction requires high fidelity throughout the processes of audiovisual acquisition, processing, and playback [23].

Overall, laboratory-based soundscape simulation has gradually developed from audio-only reproduction toward immersive systems integrating visual presentation, interaction, and spatial sound-field reconstruction. The central methodological challenge is therefore to balance experimental controllability with environmental authenticity.

This methodological framing, however, presupposes that environmental authenticity can be adequately captured by parametric indicators alone. Contemporary soundscape scholarship suggests otherwise: an acoustic environment is not merely registered but inhabited, and its qualities emerge through embodied, situated perception rather than the passive reception of acoustic signals [24,25]. Building on this phenomenological tradition, Böhme's account of atmosphere describes how the perceptual character of a place is first grasped as an undivided whole, prior to its decomposition into discrete attributes [26,27], while Schulte-Fortkamp and Jordan have argued that this concern is intrinsic, rather than external, to architectural practice [28]. Laboratory simulation therefore faces a twofold requirement: it must reproduce not only the measurable audiovisual properties of a site, but also the conditions under which a participant can become a situated listener within it. This dual requirement motivates the interactive, user-controlled design of the system developed in the present study.

1.3. The Emergence of Virtual Reality (VR) in Soundscape Research

In conventional laboratory experiments, visual stimuli are commonly presented using two-dimensional photographs, while auditory stimuli are reproduced through edited recordings captured on site. With the development of immersive technologies, since around 2010, virtual reality has increasingly been adopted in soundscape research to address some of the limitations of field studies and conventional laboratory experiments. In 2015, Preis et al. investigated audiovisual comfort across different environments by presenting field-recorded visual and auditory materials separately and in combination under laboratory conditions [16]. In 2018, Lee et al. used a laboratory-based setup combining virtual visual presentation with two-channel headphone playback to investigate the restorative effects of rural soundscapes [10]. In 2021, Lu et al. conducted audiovisual interaction experiments in urban residential environments using three-dimensional models created in Rhino and

implemented in Unity, together with simulated sound sources [29]. In 2023, Ren et al. used three-dimensional street models and a combination of site-recorded and digitally sourced sounds to investigate the influence of visible street greenery on acoustic comfort under traffic noise conditions [11].

More recently, the field has expanded rapidly. A systematic review of virtual reality applications in urban soundscape evaluation synthesized this growing body of work and identified reproduction fidelity and interaction modality as the two dominant sources of variance between studies [30]. Empirical studies have since extended the paradigm toward audiovisual biophilic interventions in indoor environments [31], longitudinal restoration effects following repeated virtual exposure [32], and higher-order Ambisonics reproduction combined with immersive visual context [33]. Parallel work has shown that visual context can systematically modulate auditory distance perception in virtual relative to real settings [34], reinforcing the need for tightly coupled audiovisual systems. Together, these developments indicate a shift from demonstrating basic feasibility toward standardizing reproduction fidelity—the concern that motivates the present study.

These studies show that virtual reality can provide controlled, repeatable, and immersive multisensory exposure environments. It can therefore be used to investigate how visual factors, sound-source information, and environmental context influence perception [35,36]. However, current VR-based soundscape studies still face two major limitations. First, the modeling of three-dimensional natural scenes requires substantial manual effort, and achieving realistic simulation effects remains difficult, particularly for dynamic natural elements such as trees, vegetation, and flowing water. Second, sound reproduction is often based on binaural recordings, making it difficult to reconstruct a realistic spatial sound field [15]. Therefore, a higher-fidelity and more interactive audiovisual simulation platform is needed.

1.4. Application of 360° Video and Spatial Audio

Owing to its relatively low cost, high level of immersion, and strong sense of realism, 360° video has become one of the most widely used forms of visual content in recent environmental perception research [37]. Compared with fully modeled three-dimensional virtual reality environments, 360° video offers three main advantages: (1) high-fidelity reproduction of real-world scenes; (2) rapid data acquisition and deployment; and (3) the ability to meet the needs of urban landscape research without complex three-dimensional modeling.

Compared with fully modeled 3D VR environments, a 360° video system can provide a panoramic visual experience, simulate movement through real-world environments, and allow users to control their viewing direction within the panoramic scene through interactive devices [38]. On the auditory side, Ambisonics can be used to reproduce sound within a three-dimensional environment [2,5]. This approach enables immersive viewing from multiple directions and supports audiovisual alignment and synchronization, thereby providing users with a stronger sense of presence [18].

Previous studies have applied 360° video to the perceptual evaluation of urban green spaces, streets, and exhibition environments [39], and recent research has combined 360° video with Ambisonics reproduction in immersive laboratory settings [33]. These developments demonstrate that audiovisual integration is already an established direction in soundscape research. However, differences in visual displays, audio reproduction methods, and interaction techniques require careful consideration when interpreting perceptual evaluations across experimental platforms [30]. Although 360° video can improve visual immersion, the development of audio systems in existing audiovisual simulation platforms often lags behind that of visual systems. Most panoramic-image-based simulation studies still rely on static panoramic images combined with conventional stereo recordings, or

support interactive visual rotation while the audio remains fixed and does not change in response to users' head movements.

High-fidelity auditory simulation is essential for virtual scene experience. The core technologies of spatial audio include (1) head-related transfer functions, whose psychophysical basis for reproducing directional and distance cues was established by Wightman and Kistler [40,41]; (2) Ambisonics, which enables omnidirectional sound-field encoding and is widely used with 360° video [42]; and (3) the simulation of early reflections and late reverberation, which enhances spatial perception and realism [43,44].

Spatial audio not only enhances immersion but also influences users' spatial cognition and behavior. Potter et al. found that spatial audio was more effective than visual clarity in enhancing immersion in virtual reality environments [21]. Hirway et al. further demonstrated through eye-tracking experiments that adding spatial audio to 360° video significantly changes the distribution of visual attention [45]. Compared with conventional stereo playback, spatial audio also provides advantages in spatial presence, sound-source localization, and perceived realism.

Previous studies have combined 360° video with Ambisonics-based spatial sound reproduction for laboratory evaluation of outdoor environments and urban soundscapes [23,33]. The present study builds on these implementations using a monitor-based platform that coordinates recorded 360° video, object-based spatial audio, and mouse-controlled viewing. Its focus is on how the mode of visual presentation and the inclusion of background sound relate to perceptual evaluations within this laboratory framework. The interactive component also provides a link to the human-centred perspectives discussed above: participants can actively explore the recorded environment rather than observe a single fixed view [46,47]. Background field recordings can be considered part of the contextual setting within which individual sound sources are experienced [48,49]. These perspectives inform the interpretation of the platform, but the experiment evaluates subjective ratings rather than directly testing embodied experience or affective anchoring. The methodological contribution therefore lies in the implementation of the coordinated audiovisual platform, while the empirical contribution concerns comparisons between presentation conditions and associations among perceptual ratings.

1.5. Research Objectives

This study develops and evaluates a laboratory platform integrating 360° video with spatial audio for built-environment perception research. The experiment compares two visual presentation modes—static photographs and 360° video—under conditions with and without background sound. Spatial audio is included in all four scenarios; therefore, the comparison does not independently isolate the contribution of spatial audio. The specific objectives of this study are as follows:

1. To implement a monitor-based audiovisual simulation platform that coordinates 360° video with spatial audio and updates the virtual listener orientation according to mouse-controlled viewing.
2. To compare ratings of visual perception, auditory perception, spatial perception, perceived realism, and overall perception between static photographs and 360° video under each background-sound condition, and to examine differences between background-sound conditions within each visual presentation mode.
3. To examine whether differences in perceived audiovisual characteristics between the 360° video and photograph conditions are associated with differences in scene-quality ratings.

2. System Implementation and Experimental Design

2.1. Implementation of the Laboratory-Based 360° Video and Spatial Audio Simulation System

A 360° video is generated by synchronously capturing the same environment from multiple viewpoints using several wide-angle camera lenses. The recorded images are stitched into a panoramic image and mapped onto a spherical coordinate system. During playback, a virtual camera displays the corresponding region of the spherical image according to the viewing angle and direction, thereby presenting the visual scene. However, conventional 360° video playback has a limitation during interactive control. When the video is paused or played at a faster or slower speed, the accompanying audio must also be paused or adjusted by the same proportion, making it difficult to achieve fully interactive audiovisual playback.

To address this limitation, this study developed a simulation system that dynamically integrates 360° video with a spatial audio scene. The 360° video is used for visual presentation, while spatial audio technology is used to construct a three-dimensional sound field corresponding to the visual scene. During video playback, the position represented in the panoramic video is mapped to the corresponding position in the spatial audio scene. Using information about the user's position and viewing direction, the system synchronizes the playback parameters of the 360° video and spatial audio, thereby providing an integrated audiovisual experience.

For the implementation of the spatial audio component, audio-recording equipment was used to collect sound signals as source materials for subsequent system development. Along the predefined experimental route, various environmental sounds were recorded on site, including birdsong, cicada chirping, wind sounds, insect sounds, human speech, and traffic noise. The recordings were then processed in Adobe Audition to remove unwanted noise and extract individual sound-source clips.

A three-dimensional audio scene corresponding to the spatial environment was subsequently constructed. The processed recordings were treated as sound sources in the reconstructed sound field, and a listener object was positioned at the visual observation point. Based on the locations of the sound sources, the environmental settings, and the listener's position, a three-dimensional audio engine was used to simulate and reproduce the sound field in real-time. This process enabled synchronized visual and auditory presentation and provided users with a stronger sense of presence during interaction.

Full wave-based acoustic simulation in virtual built environments involves substantial computational cost and is therefore difficult to perform in real-time [29]. Simplified methods are commonly adopted in three-dimensional games and multimedia applications. These methods primarily simulate sound direction, distance, and spatial characteristics by adjusting the relative gains between audio channels, applying head-related transfer functions to reproduce directional and positional cues, modifying sound level and frequency response according to the distance between the sound source and the listener, and simulating early reflections and late reverberation to enhance spatial perception and realism [50]. Compared with indoor acoustic environments, outdoor and urban landscape environments generally have less stringent requirements for acoustic simulation accuracy. In this study, spatial audio was implemented in Unreal Engine through the IAudioSpatialization interface. The system used object-based HRTF spatialization for individual sound sources, rather than Ambisonics B-format decoding. The renderer generated binaural headphone output according to the positions of the sound sources relative to the virtual listener. The listener orientation was updated according to the mouse-controlled viewing direction. A network transmission protocol supported the exchange of position information and playback-control commands. The overall system construction process is shown in Figure 1.

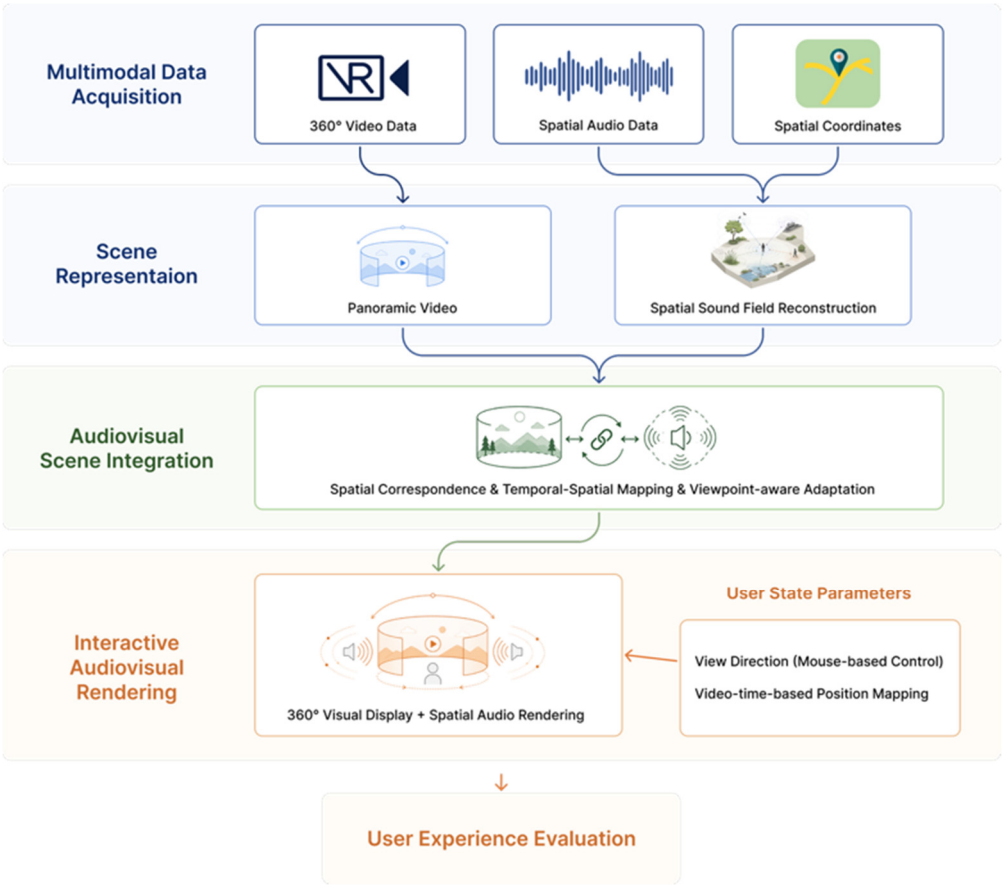


Figure 1. Construction process of the proposed audiovisual simulation system. (Arrows indicate the flow of data and control information, while different colors distinguish the stages of the system workflow).

Compared with conventional approaches, the proposed 360° video and spatial audio system offers several advantages:

- First, it improves production efficiency. A panoramic camera mounted on a simple head-supported rig can be used to collect data while the operator walks along the predefined route. This procedure is relatively simple and reduces the time and labor required for environmental data acquisition.
- Second, because 360° video is essentially a video stream, its playback generally requires fewer graphics-computing resources than fully modeled virtual reality environments. It therefore has lower hardware requirements and can be more easily applied in a wider range of experimental settings.
- Third, the system provides an interactive method for experiencing recorded environments. Operations of the 360° video player are mapped to physical actions such as walking and turning on a walking simulator. Route selection and positional transitions on the map are also mapped to corresponding video navigation operations. These mappings enable users to interact with the recorded scene and provide a distinctive form of virtual environmental experience.

2.2. Experimental Design

The plaza in front of the Red Building (Building No. 6) on the Wushan Campus of South China University of Technology was selected as the experimental site. The area contains historic buildings associated with the former Sun Yat-sen University campus and is characterized by dense and diverse vegetation, an attractive landscape, and abundant soundscape resources.

Five representative soundscape experience and evaluation nodes were selected at relatively even intervals along a walking route around the plaza, as indicated by the red circles in Figure 2. The route passed sequentially through Nodes 1–5, following the yellow dashed line shown in Figure 2. Video footage was recorded along the predefined route and subsequently integrated into the platform to construct the laboratory-based simulation system.

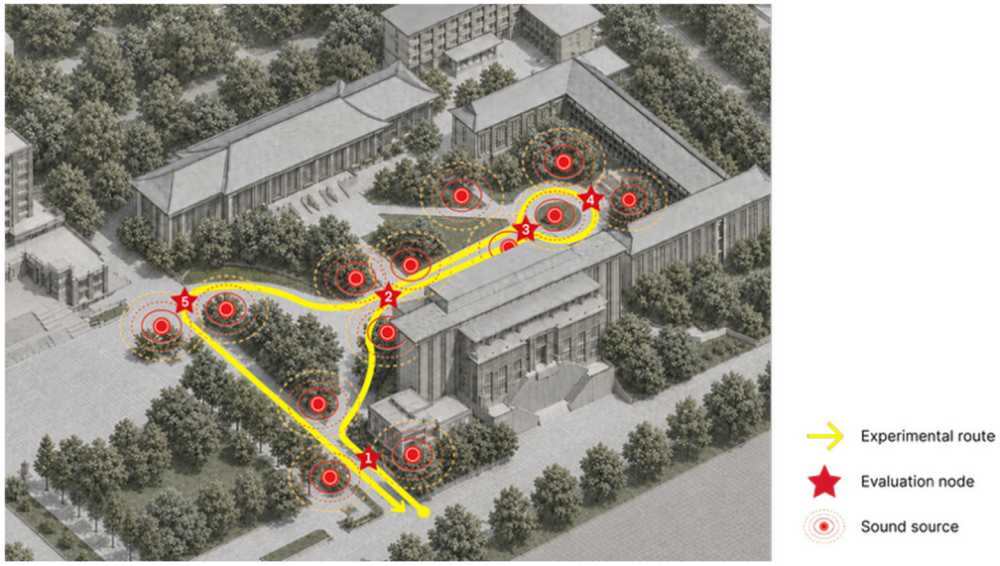


Figure 2. Experimental route and locations of the five evaluation nodes. (Numbers 1–5 indicate the five evaluation nodes).

2.2.1. Visual Stimuli

The selected site contained a variety of visual elements, including campus buildings, green spaces, roads, sculptures, and vegetation. The five nodes were distributed relatively evenly along the walking route. This arrangement ensured that the dynamic video and audio presented before and after each node had a playback duration of approximately 40 s.

Panoramic videos were recorded using a 360° panoramic camera, the Insta360 ONE X2, as shown in Figure 3. To capture the surrounding environment as completely as possible, the camera was mounted above the operator's head using a head-supported rig during recording. A GPS device was also used to record the current position, and the resulting location information was used for spatial positioning and navigation during playback.



Figure 3. Panoramic video recording using the Insta360 ONE X2.

For the static visual condition, two-dimensional images extracted from the panoramic video at each node were used as visual stimuli. For the dynamic visual condition, a customized 360° video player was developed to provide a playback-control interface. The video was presented on a computer and matched with the virtual sound-field model created in Unreal Editor, as shown in Figure 4. The participant's current position was synchronized with the virtual environment to create the experience of walking through the scene.



Figure 4. Virtual sound-field modeling in Unreal Editor.

During the simulated walking experience, the playback time of the panoramic video was mapped to the participant's position in the virtual audio scene. During interaction in the 360° video condition, the virtual viewing direction was controlled using a mouse. The resulting orientation data were transmitted to the audio engine, which calculated the spatial audio rendering in real-time. The generated binaural spatial audio was then delivered to the participant through headphones, thereby achieving synchronized visual and auditory presentation. Table 1 presents screenshots of the dynamic 360° video stimuli at the five nodes.

Table 1. Visual materials.

	Location 1	**Location 2**	**Location 3**	**Location 4**	**Location 5**
Photo	P1	P2	P3	P4	P5
Panoramic Video	Front	Front	Front	Front	Front
	Right	Right	Right	Right	Right
	Back	Back	Back	Back	Back
	Left	Left	Left	Left	Left

2.2.2. Collection of Auditory Materials and Construction of the Soundscape

Background sound was recorded simultaneously with the video using a Sony PCM-D100 portable recorder. The total recording duration was 4 min 55 s, and the recordings were used as background-sound materials in the subsequent experiment. During the formal experiment, the corresponding 40 s audio segment was played at each node to ensure a consistent listening duration across the different conditions.

Because the occurrence of natural sounds could not be controlled during video recording, additional in situ audio recordings were conducted along the predefined walking route and within the experimental site based on GPS positioning. The sounds were recorded using Sennheiser AMBEO binaural recording headphones. The recordings were then processed in Adobe Audition to remove unwanted noise, extract natural sound sources, and classify the resulting materials. A total of 15 types of natural sound materials were obtained, as shown in Table 2, and were subsequently used as source materials for system development.

Table 2. Auditory materials.

Cicada Chirping Material 1	Cicada Chirping Material 2	Insect Chirping Material 1
Insect Chirping Material 2	Insect Chirping Material 3	Wind Sound Material 1
Wind Sound Material 2	Bird Chirping Material 1	Bird Chirping Material 2
Bird Chirping Material 3	Bird Chirping Material 4	Bird Chirping Material 5
Bird Chirping Material 6	Bird Chirping Material 7	Bird Chirping Material 8

A virtual three-dimensional spatial model was constructed in Unreal. The site map was integrated into the model and spatially referenced according to the GPS location data. The natural sound recordings were then arranged as three-dimensional sound sources within

the virtual environment in a manner intended to approximate the actual site conditions. For example, insect sounds were attached to lawn areas, while birdsong was positioned in trees, as shown in Table 3.

Table 3. Auditory element design at node locations.

Sound Category	Location 1	Location 2	Location 3	Location 4	Location 5
Virtual Sound	Birdsong 1 and Birdsong 3	Birdsong 2 and Birdsong 3	Birdsong 4 and Birdsong 5	Birdsong 6 and Birdsong 7	Birdsong 8 and Birdsong 4
	Insect chirping 1	Wind sound 5	Wind sound 6	Wind sound 6	Insect chirping 3
	Wind sound 3	Cicada chirping 1	Insect chirping 2 and Insect chirping 3	Insect chirping 2 and Insect chirping 3	
			Cicada chirping 2	Cicada chirping 1	
Background Sound	Background Sound 1	Background Sound 2	Background Sound 3	Background Sound 4	Background Sound 5

Based on this spatial arrangement, the positions and types of the sound sources, together with the relevant scene parameters, were further adjusted. These parameters included propagation range, gain, and attenuation. A degree of randomness was introduced into the playback timing and loudness adjustment to produce a more natural effect that was closer to the conditions of the actual site.

Sound-field reconstruction materials were generated for the corresponding environmental conditions. During auditory playback, five soundscape configurations corresponding to the five nodes were created. These configurations were divided into two conditions: with background sound and without background sound.

A binaural artificial head and a sound-level calibrator were used to calibrate the sound pressure levels of the two auditory conditions. Because the sound sources fluctuated over time, the equivalent continuous A-weighted sound pressure level was controlled across the entire playback period. The sound pressure levels in both conditions ranged from 43 to 65 dB (A).

3. Experimental Procedure

3.1. Participants

A total of 38 participants (16 males and 22 females) participated in the experiment. All participants were undergraduate or graduate students from the School of Architecture at South China University of Technology (Wushan Campus). Strict inclusion criteria were applied during participant recruitment, and all participants self-reported normal hearing and normal vision. Familiarity with the study site was assessed using the visit-frequency categories in Questionnaire 1. Among the 36 participants included in the final analysis, 7 (19.4%) were classified as unfamiliar, 15 (41.7%) as somewhat unfamiliar, 9 (25.0%) as moderately familiar, 2 (5.6%) as somewhat familiar, and 3 (8.3%) as familiar.

3.2. Experimental Design and Materials

To investigate whether the proposed 360° video and 3D audio audiovisual simulation system provides advantages over the traditional 2D photograph-based evaluation method, four experimental scenarios were designed by combining different visual presentation modes and background sound conditions.

The experiment included two visual presentation modes: a traditional 2D photograph condition and a 360° video condition. The 2D photograph condition presented static images

of the environment, while the 360° video condition provided panoramic visual information that allowed participants to explore the environment through viewpoint changes. In addition, two auditory conditions were included: with background sound and without background sound.

Based on these two factors, four experimental scenarios were developed, as summarized in Table 4:

Scenario 1: 2D photograph + spatial audio without background sound;
Scenario 2: 2D photograph + spatial audio with background sound;
Scenario 3: 360° video + spatial audio without background sound;
Scenario 4: 360° video + spatial audio with background sound.

Table 4. Types of combinations of visual and auditory elements in the audiovisual materials.

	Visual	Auditory
Scenario 1	2D static photo	Virtual sound recording playback
Scenario 2	2D static photo	Virtual environmental sound recording playback + Background sound
Scenario 3	VR360	Virtual environmental sound recording playback
Scenario 4	VR360	Virtual environmental sound recording playback + Background sound

To compare the effects of different visual presentation modes under the same auditory condition, the four scenarios were divided into two comparison groups, as illustrated in Figure 5.

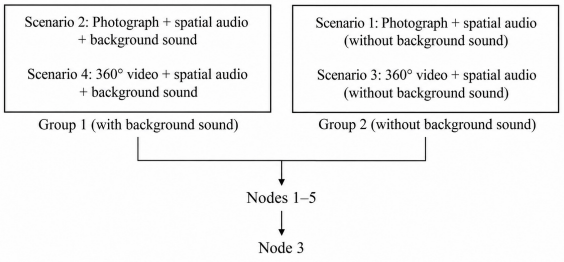


Figure 5. Experimental data grouping.

Group 1 compared Scenario 2 and Scenario 4, both of which included background sound. This comparison examined the differences between the traditional 2D photograph condition and the proposed 360° video condition when background sound was provided.

Group 2 compared Scenario 1 and Scenario 3, both of which were presented without background sound. This comparison examined the differences between the two visual presentation modes in the absence of background sound.

In addition, comparisons between Scenario 1 and Scenario 2, and between Scenario 3 and Scenario 4, were conducted to further examine the effects of background sound under the same visual presentation mode.

Each scenario contained five spatial nodes corresponding to different viewpoints within the experimental environment. Participants experienced the audiovisual materials of each node sequentially to evaluate their perceptions of the simulated environment.

3.3. Experimental Setup and Environment

The experiment was conducted in a well-soundproofed indoor laboratory. The experimental system consisted of a computer equipped with an NVIDIA RTX 4090 GPU, which

was responsible for processing and rendering the visual, auditory, and integrated 360° video and spatial audio data used in the experiment, as shown in Figure 6.



Figure 6. Experimental setup and equipment.

Visual stimuli were presented on a 60-inch monitor. All participants, including the two who discontinued the experiment because of dizziness, used a mouse to control the virtual viewing direction in the 360° video condition. No head-mounted display or head-tracking device was used. Mouse input controlled rotation of the virtual view.

The auditory stimuli were played through the computer's sound card and delivered using Sennheiser headphones. This setup enabled participants to perceive spatial auditory information, including sound direction and distance.

During interaction in the 360° video condition, mouse-controlled changes in the virtual viewing direction were transmitted to the spatial audio engine in real-time. Based on the orientation information, the system calculated the corresponding binaural spatial audio rendering and updated the sound field accordingly. This process enabled the spatial audio presentation to respond dynamically to changes in viewing direction while maintaining synchronization between the visual and auditory information.

3.4. Experimental Procedure

Before the experiment, participants were provided with a participant information sheet explaining the research objectives, experimental procedures, potential risks, and data usage. All participants were required to sign an informed consent form before participating in the experiment. Before the formal experiment, participants wore headphones and completed a 5-min training session using the mouse to practice viewpoint control and node switching.

A calibration procedure was then conducted to adjust the display brightness and headphone volume according to each participant's preferred comfort level. After confirming that they did not experience any physical discomfort, participants completed Questionnaire 1 and proceeded with the formal experiment. During the experiment, participants experienced the four experimental scenarios sequentially. Within each scenario, the audiovisual materials corresponding to the five spatial nodes were presented in sequence, with a 10 s interval between adjacent nodes. After completing each scenario, participants were allowed a 5-min rest period and then completed Questionnaire 2 to evaluate their experience of that scenario.

3.5. Experimental Design and Analysis

3.5.1. Questionnaires and Variables

Two questionnaires were used in the experiment and were completed by the participants on a tablet device. Questionnaire 1 was completed before the formal experiment. It contained seven questions covering gender, age group, educational attainment, academic major, frequency of visits to the study site, self-reported hearing ability, and participant identification number. A blank version of Questionnaire 1 is provided in the Supplementary Materials as Questionnaire S1.

Questionnaire 2 was completed after each experimental scenario. It contained eight numbered questions recording the participant identification number, scenario and node, scene-quality ratings, the ranking of sound sources by perceived prominence, ratings of natural-sound loudness and satisfaction, and perceived audiovisual characteristics.

The present study focused on the five scene-quality items in Question 4 and the nine semantic-differential items in Question 8. The scene-quality items assessed visual perception, auditory perception, spatial perception, perceived realism, and overall perception on a five-point scale ranging from 1 (very dissatisfied) to 5 (very satisfied). The audiovisual-characteristic items used five-point scales between nine pairs of bipolar descriptors: blurred–clear, artificial–natural, monotonous–rich, intermittent–continuous, annoying–pleasant, single-directional–multi-directional, enclosed–open, conflicting–harmonious, and detached–immersive.

The five scene-quality indicators were treated as outcome variables for evaluating differences between the proposed 360° video and spatial audio system and the conventional 2D photograph-based soundscape evaluation method. The nine semantic-differential indicators were treated as perceived audiovisual-characteristic variables and were used to examine which perceived characteristics were associated with differences in scene-quality evaluations.

3.5.2. Data Collection and Processing

Data from 36 participants were retained for the final analysis after the exclusion of two participants who reported dizziness during the experiment. The questionnaire responses for the four experimental scenarios were checked and organized according to the two comparison groups defined in Section 3.2.

For Group 1, paired data from Scenario 2 and Scenario 4 were used to compare the 2D photograph and 360° video conditions with background sound. For Group 2, paired data from Scenario 1 and Scenario 3 were used to compare the two visual presentation modes without background sound.

For the subsequent comparative and correlation analyses, difference scores were calculated between the 360° video condition and the corresponding 2D photograph condition. The difference score for Group 1 was calculated as Scenario 4 minus Scenario 2, while that for Group 2 was calculated as Scenario 3 minus Scenario 1.

Difference scores were calculated separately for the five scene-quality indicators and the nine perceived audiovisual-characteristic indicators. Node 3 was selected as a common reference point for comparing the experimental scenarios because it occupied the midpoint of the five-node sequence within each scenario. The rationale was to use an intermediate observation point, allowing participants some experience of the current scenario before evaluation while avoiding the end of the node sequence. The comparative and correlation analyses used node-specific ratings from Node 3 rather than ratings aggregated across all five nodes.

For the descriptive spatial maps, overall-perception ratings were summarized separately for each scenario and each of the five nodes. Where duplicate records shared the same participant identifier within a scenario–node combination, their ratings were averaged before calculating the group mean. Mean ratings were displayed at the corresponding locations using a common 1–5 color scale, without spatial interpolation.

3.5.3. Analytical Methods

Differences in scene-quality ratings and perceived audiovisual-characteristic ratings between the paired experimental scenarios were examined using the Wilcoxon signed-rank test [51]. Effect sizes for the Wilcoxon signed-rank tests were calculated as $r = |Z| / \sqrt{N}$, where N denotes the number of complete participant pairs in each comparison (N = 36). To

examine the associations between perceived audiovisual characteristics and scene-quality evaluations, Spearman's rank correlation analysis was conducted [52]. For each comparison group, the first set of variables consisted of the difference scores for the five scene-quality indicators: visual perception, auditory perception, spatial perception, perceived realism, and overall perception.

The second set of variables consisted of the difference scores for the nine perceived audiovisual-characteristic indicators. These scores represented the participants' perceived differences between the 360° video condition and the corresponding 2D photograph condition.

Spearman's rank correlation analysis was used to examine the associations between changes in perceived audiovisual characteristics and changes in scene-quality ratings. All statistical analyses were performed using IBM SPSS Statistics, version 22.0.

4. Results

4.1. Scores Across the Four Experimental Scenarios

Table 5 presents the means (M) and standard deviations (SD) medians, and interquartile ranges (IQRs) of the scene-quality and perceived audiovisual-characteristic ratings across the four experimental scenarios. The mean scores for the five scene-quality indicators ranged mainly from 3.17 to 4.19. Scenario 4 received the highest mean ratings for visual perception and spatial perception and relatively high ratings for perceived realism and overall perception. Scenario 3 received the highest mean rating for auditory perception. Across the four scenarios, the standard deviations of the scene-quality ratings ranged from 0.599 to 1.017. Overall-perception ratings showed relatively low variability, whereas auditory-perception ratings showed comparatively greater variability among participants.

Table 5. Means (M) and standard deviations (SD), medians, and interquartile ranges (IQRs) of the scene-quality and perceived audiovisual-characteristic ratings across the four experimental scenarios.

Indicator	Scenario 2				Scenario 1				Scenario 4				Scenario 3			
	M	SD	Median	IQR	M	SD	Median	IQR	M	SD	Median	IQR	M	SD	Median	IQR
Scene-quality ratings																
Visual perception	3.25	0.874	3.00	1.00	3.17	0.775	3.00	1.00	4.19	0.710	4.00	1.00	4.11	0.708	4.00	1.00
Auditory perception	3.22	1.017	3.00	2.00	3.69	0.856	4.00	1.00	3.69	0.786	4.00	1.00	3.86	0.683	4.00	0.75
Spatial perception	3.25	0.937	3.00	1.00	3.56	0.809	3.00	1.00	4.19	0.668	4.00	1.00	3.81	0.710	4.00	1.00
Perceived realism	3.33	0.793	3.00	1.00	3.22	0.832	3.00	1.00	3.92	0.732	4.00	1.00	3.64	0.762	4.00	1.00
Overall perception	3.17	0.775	3.00	1.00	3.39	0.599	3.00	1.00	3.94	0.754	4.00	0.75	3.75	0.692	4.00	1.00
Perceived audiovisual-characteristic ratings																
Blurred–Clear	3.69	0.951	4.00	1.00	3.83	0.878	4.00	1.00	4.08	0.604	4.00	0.00	4.08	0.604	4.00	0.00
Artificial–Natural	3.08	0.996	3.00	1.75	3.72	1.031	4.00	0.00	3.22	0.898	3.00	1.00	3.97	0.774	4.00	1.50
Monotonous–Rich	3.58	0.906	4.00	1.00	3.28	1.186	3.50	2.00	3.83	0.878	4.00	1.00	3.81	0.786	4.00	1.00
Intermittent–Continuous	3.22	0.929	3.00	2.00	3.03	1.055	3.00	2.00	3.64	0.931	4.00	1.00	3.14	1.018	3.00	2.00
Annoying–Pleasant	3.08	0.874	3.00	1.00	3.56	0.809	4.00	1.00	3.36	0.683	3.00	1.00	3.89	0.575	4.00	0.00
Single–Diverse	3.58	1.105	4.00	1.00	3.28	0.914	3.00	1.00	3.78	0.959	4.00	1.00	3.64	0.723	4.00	1.00
Closed–Open	3.56	0.909	3.00	1.00	3.61	0.934	4.00	1.00	3.81	0.920	4.00	0.75	3.78	0.797	4.00	1.00
Conflicting–Harmonious	2.97	1.055	3.00	2.00	3.47	0.910	4.00	1.00	3.75	0.692	4.00	1.00	3.89	0.820	4.00	1.00
Detached–Immersive	3.00	1.042	3.00	2.00	3.17	0.737	3.00	0.75	3.56	0.939	4.00	1.00	3.75	0.806	4.00	1.00

The mean scores for the nine perceived audiovisual-characteristic indicators ranged from 2.97 to 4.08. Scenarios 4 and 3 received relatively high mean ratings across several indicators. Scenario 3 received the highest ratings for monotonous–rich and intermittent–continuous, whereas Scenario 3 received the highest ratings for artificial–natural, annoying–

pleasant, conflicting–harmonious, and detached–immersive. Both Scenarios 4 and 3 received the same highest mean rating for blurred–clear.

The standard deviations of the perceived audiovisual-characteristic ratings ranged from 0.575 to 1.186. Ratings for indicators such as artificial–natural and monotonous–rich showed relatively greater variability among participants, whereas ratings for blurred–clear showed comparatively lower variability.

4.2. Differences in Scene Quality Evaluation

Figures 7 and 8 present the perceived scores of the two comparison groups across different evaluation indicators. For Group 1 (Scenario 2 vs. Scenario 4, with background sound), the results showed that compared with the 2D photograph-based evaluation condition, the 360° video condition received significantly higher ratings for visual perception ($r = 0.693$, $Z = 4.156$, $p < 0.001$), spatial perception ($r = 0.707$, $Z = 4.242$, $p < 0.001$), perceived realism ($r = 0.514$, $Z = 3.084$, $p = 0.002$), and overall perception ($r = 0.668$, $Z = 4.007$, $p < 0.001$). Auditory perception also showed a significant difference between the two conditions ($r = 0.419$, $Z = 2.514$, $p = 0.012$).

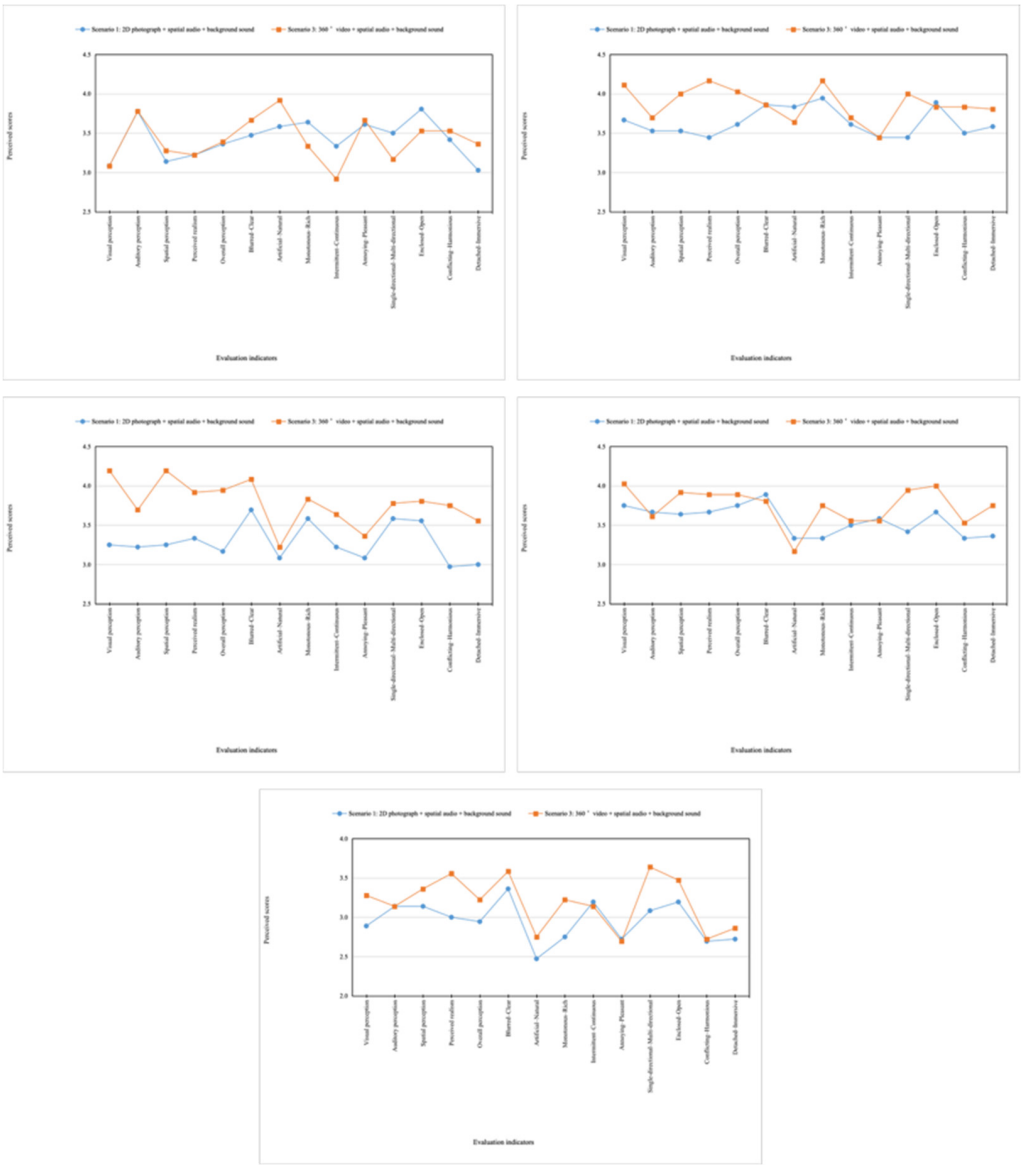


Figure 7. Perceived scores for different evaluation metrics (Scenario 2 vs. Scenario 4).

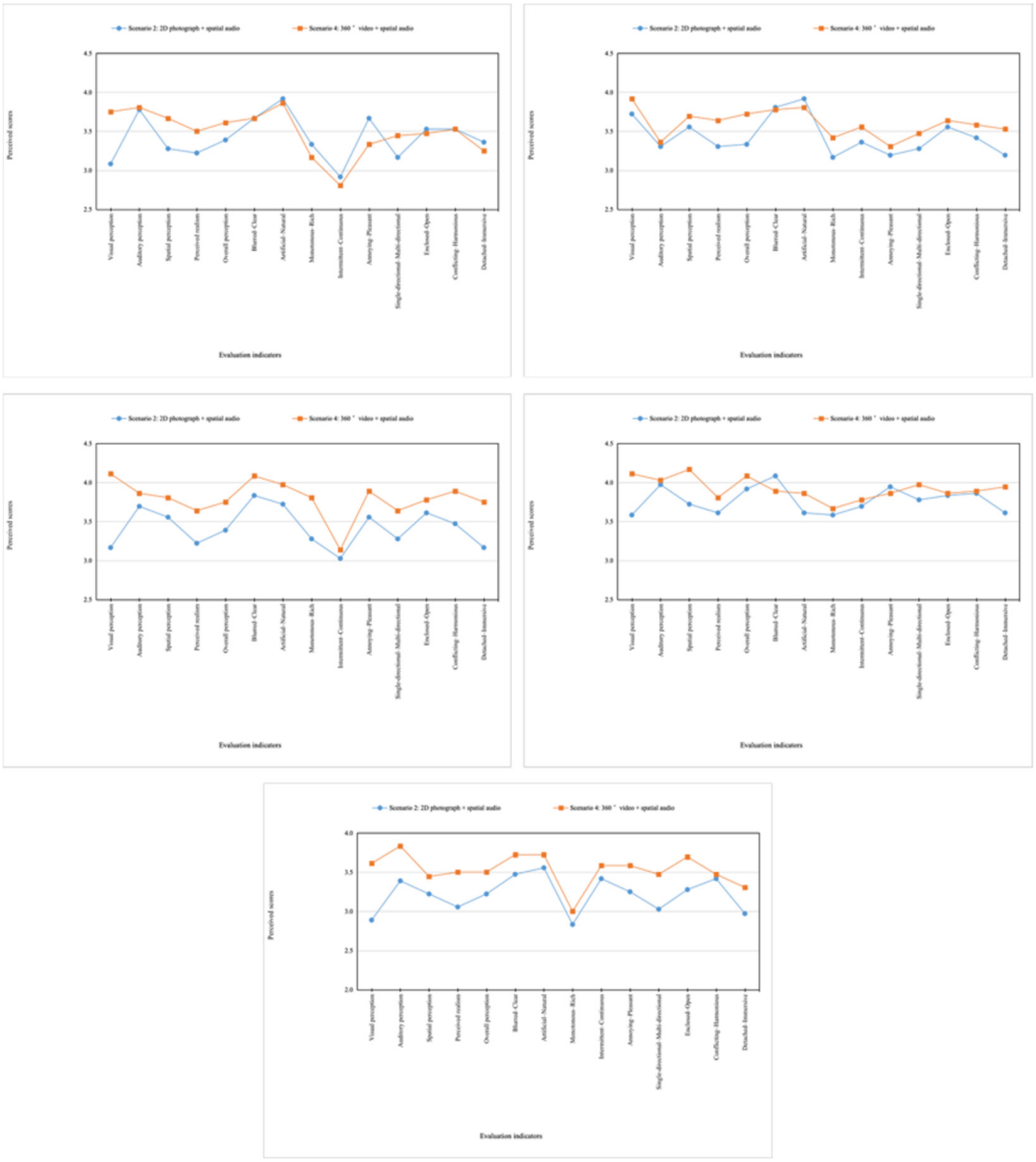


Figure 8. Perceived scores for different evaluation metrics (Scenario 1 vs. Scenario 3).

For Group 2 (Scenario 1 vs. Scenario 3, without background sound), the results showed that compared with the 2D photograph-based evaluation condition, the 360° video condition received significantly higher ratings for visual perception (r = 0.672, Z = 4.034, $p < 0.001$), perceived realism (r = 0.414, Z = 2.486, $p = 0.013$), and overall perception (r = 0.379, Z = 2.272, $p = 0.023$). No significant differences were observed for auditory perception (r = 0.175, Z = 1.048, $p = 0.295$) or spatial perception (r = 0.287, Z = 1.719, $p = 0.086$).

To further examine the role of background sound, additional paired comparisons were conducted between Scenario 1 and Scenario 2 under the 2D photograph condition, and between Scenario 3 and Scenario 4 under the 360° video condition.

Under the 2D photograph condition, Scenario 2 showed significantly higher auditory-perception ratings than Scenario 1 (r = 0.426, Z = 2.555, $p = 0.011$). Under the 360° video condition, a significant difference was observed in spatial-perception ratings between Scenario 3 and Scenario 4 (r = 0.456, Z = 2.738, $p = 0.006$). These findings suggest that background sound may contribute differently to scene-quality evaluations depending on the visual presentation mode.

4.3. Associations Among Perceptual Ratings and Their Spatial Distribution

Figure 9 presents the Spearman rank correlation analysis between differences in scene-quality ratings and differences in perceived audiovisual-characteristic ratings under conditions without and with background sound. Under the condition without background sound, significant positive associations were observed between several scene-quality indicators and perceived audiovisual characteristics. Visual perception showed moderate positive correlations with clarity ($r_s = 0.526$, $p = 0.001$) and richness ($r_s = 0.411$, $p = 0.013$). Auditory perception was significantly associated with perceived naturalness ($r_s = 0.448$, $p = 0.006$), continuity ($r_s = 0.391$, $p = 0.018$), pleasantness ($r_s = 0.544$, $p = 0.001$), harmony ($r_s = 0.562$, $p < 0.001$), and immersion ($r_s = 0.374$, $p = 0.025$). Changes in spatial perception were moderately correlated with clarity ($r_s = 0.466$, $p = 0.004$), richness ($r_s = 0.333$, $p = 0.047$), and openness ($r_s = 0.521$, $p = 0.001$). Perceived realism was significantly influenced by clarity ($r_s = 0.484$, $p = 0.003$), richness ($r_s = 0.492$, $p = 0.002$), pleasantness ($r_s = 0.396$, $p = 0.017$), multidirectionality ($r_s = 0.372$, $p = 0.026$), and openness ($r_s = 0.490$, $p = 0.002$). Improvements in overall scene quality were associated with perceived richness ($r_s = 0.447$, $p = 0.006$), pleasantness ($r_s = 0.516$, $p = 0.001$), multidirectionality ($r_s = 0.396$, $p = 0.017$), openness ($r_s = 0.337$, $p = 0.044$), harmony ($r_s = 0.405$, $p = 0.014$), and immersion ($r_s = 0.330$, $p = 0.049$).

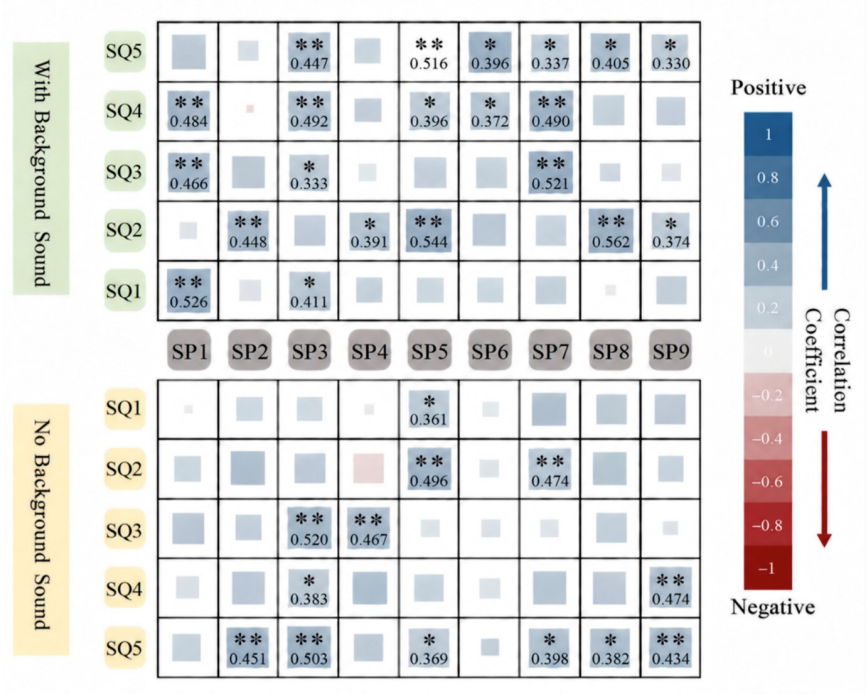


Figure 9. Spearman rank correlation analysis of scene quality (SQ) and scene feature evaluation (SP) with and without background sound (N = 36). (Note: * Correlation is significant at the 0.05 level (2-tailed), and ** Correlation is significant at the 0.01 level (2-tailed)).

Under the background-sound condition, visual perception showed a moderate positive correlation with pleasantness ($r_s = 0.361$, $p = 0.031$). Auditory perception was significantly associated with pleasantness ($r_s = 0.496$, $p = 0.002$) and openness ($r_s = 0.474$, $p = 0.003$). Spatial perception was positively correlated with richness ($r_s = 0.520$, $p = 0.001$) and continuity ($r_s = 0.474$, $p = 0.004$). Perceived realism was mainly influenced by richness ($r_s = 0.383$, $p = 0.021$) and immersion ($r_s = 0.474$, $p = 0.004$). Overall scene quality was significantly associated with perceived naturalness ($r_s = 0.451$, $p = 0.006$), richness ($r_s = 0.503$, $p = 0.002$), pleasantness ($r_s = 0.369$, $p = 0.027$), openness ($r_s = 0.398$, $p = 0.016$), harmony ($r_s = 0.382$, $p = 0.022$), and immersion ($r_s = 0.434$, $p = 0.008$).

To complement the correlation analysis at Node 3, Figure 10 presents a descriptive spatial overview of the mean overall-perception ratings across all five evaluation nodes under the four experimental scenarios. Ratings are mapped onto the corresponding node locations using a common color scale. This visualization links the questionnaire results to their spatial context and allows for a comparison of the observed rating patterns across locations and presentation conditions. The maps are descriptive and do not establish statistically significant differences between nodes.

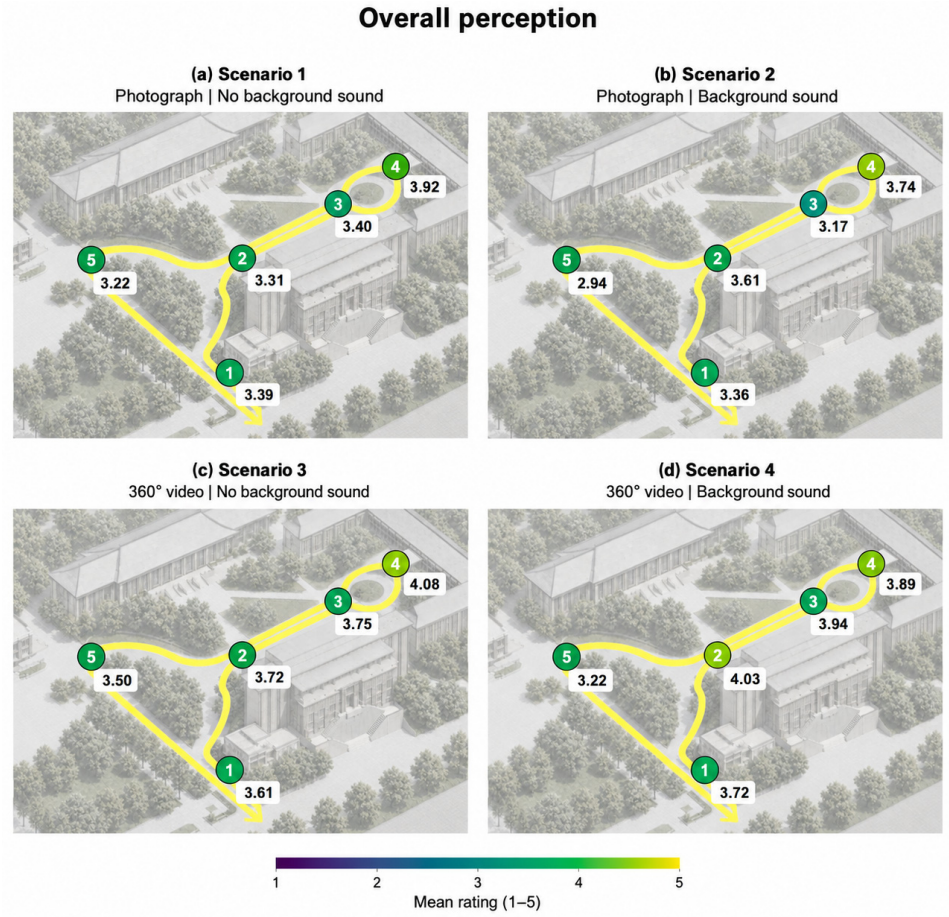


Figure 10. Spatial distribution of the mean overall-perception ratings across the five evaluation nodes under the four experimental scenarios. (Yellow arrows indicate the experimental route. Numbered circles identify the five evaluation nodes, and the adjacent values indicate the mean overall-perception ratings).

5. Discussion

5.1. Perceptual Evaluation Under Static Photograph and 360° Video Conditions

The higher ratings of visual perception, perceived realism, and overall perception in the 360° video conditions suggest that visual presentation shaped the participants' evaluations within the audiovisual framework. Compared with static photographs, 360° video provides temporal information and opportunities to explore the surrounding scene through changes in viewing direction. These features may help participants understand the environmental context and account for the observed rating differences. Because temporal information, panoramic coverage, and viewpoint control varied together, the findings concern the presentation conditions as a whole.

The opportunity to select and explore views also connects the framework to the active and situated understanding of perception discussed in Section 1.2. Participants could direct their attention within the recorded environment, while background sound provided an

auditory context for individual sound sources. This offers a way to interpret the experience beyond differences in image presentation alone, although mouse-controlled exploration represents only a limited form of engagement with the environment.

The findings complement previous research while addressing a different experimental comparison. Potter et al. examined the relative contributions of visual and spatial-audio rendering to immersion [21]. Here, spatial audio was retained across all conditions, and the experiment examined visual presentation under two background-sound conditions. Georgiou et al. emphasized fidelity throughout audiovisual recording, processing, and reproduction when replicating outdoor environments using VR and Ambisonics [23]. The present monitor-based framework instead examines perceptual evaluation using headphone playback and mouse-controlled viewing. Differences in reproduction and interaction methods should therefore be considered when relating the findings across studies.

The contribution of this comparison is to show how perceptual ratings differ between static and interactive panoramic presentation within a common audiovisual framework. Higher perceived-realism ratings reflect the participants' judgments of the presentation; establishing how closely these judgments correspond to experience in the physical environment requires matched in situ evaluation.

5.2. Different Perceptual Factors Associated with Scene Quality Evaluation

The correlation analysis showed that different audiovisual characteristics were associated with different aspects of scene-quality evaluation. These results indicate that scene quality in immersive audiovisual environments is not determined by a single perceptual factor, but by the combined contribution of multiple environmental characteristics.

Across both background-sound conditions, richness and pleasantness showed relatively consistent positive associations with overall perception. Richness may represent the amount and diversity of environmental information perceived by the participants, including visual details and auditory variations. A richer audiovisual environment may provide more spatial cues and contribute to a stronger sense of presence. Meanwhile, pleasantness reflects the emotional response generated by the environment, which may influence the participants' overall evaluation of the experience. These findings suggest that both informational richness and positive emotional responses are important aspects of immersive environmental evaluation.

Other audiovisual characteristics showed more context-dependent associations. Under the condition without background sound, clarity was positively associated with visual perception, spatial perception, and perceived realism. This suggests that when auditory information was limited, participants may have relied more heavily on visual details when evaluating the environment. However, these associations were not significant after background sound was introduced. One possible explanation is that additional auditory information may encourage participants to integrate multiple sensory cues rather than relying primarily on visual clarity.

Openness and continuity were also associated with specific perceptual dimensions. Openness showed positive associations with spatial perception and overall perception under different conditions, suggesting that perceived spatial extent may contribute to environmental experience. Continuity was associated with spatial perception when background sound was present, indicating that coherent audiovisual information may support the participants' understanding of environmental relationships. However, these associations should be interpreted as relationships between perceptual evaluations rather than direct causal effects.

Overall, the correlation results distinguish characteristics consistently associated with overall perception from those whose associations depend on the evaluation dimension and

background-sound condition. Richness and pleasantness were positively associated with overall perception under both background-sound conditions, making them relevant indicators for evaluating the overall audiovisual experience. Clarity, openness, continuity, and immersion should be examined in relation to specific perceptual outcomes and sound conditions. These findings support an evaluation approach that combines indicators relevant across conditions with a condition-specific assessment of other perceived characteristics.

5.3. Implications for Built-Environment Evaluation

The proposed framework provides a controlled setting for examining how visual presentation and auditory context relate to the perception of recorded built environments. A potential application is the evaluation of public-space nodes and locations along pedestrian routes, where vegetation, enclosure, visibility, and sound sources form different audiovisual settings. Presenting these locations through a consistent protocol could help researchers and designers compare users' experiences while retaining the surrounding visual and auditory context.

The findings also inform the selection of perceptual indicators. Differences in perceived richness and pleasantness were positively associated with differences in overall perception under both background-sound conditions. These characteristics therefore warrant attention alongside overall-perception ratings. For evaluations focused on visual perception, auditory perception, spatial perception, or perceived realism, additional indicators should be selected according to the evaluation objective and examined within the relevant sound condition. This approach uses the observed associations to guide assessment rather than treating individual characteristics as direct design prescriptions.

For site-based applications, ratings from multiple nodes could be linked to a site plan and considered alongside information on landscape composition, spatial configuration, and sound-source distribution. Such comparisons could help identify locations for closer investigation, including seating areas, entrances, or transitions along pedestrian routes. Combined with field measurements and behavioral observations, the framework could provide supplementary evidence for public-space assessment and discussions of environmental improvements. Its immediate value lies in supporting structured comparisons of recorded settings; evaluating proposed physical changes would require corresponding visual and auditory representations.

5.4. Limitations and Future Research

The findings should be interpreted within the scope of the sample, setting, and experimental design. Participants were recruited from a university population, and the stimuli represented a single campus site. Although site familiarity was documented, its relationship with perceptual ratings was not examined. Further studies should include more diverse participants and urban environments and assess whether familiarity and disciplinary background are associated with evaluation outcomes. Matched laboratory and in situ assessments would also help establish how well the framework preserves perceptual differences observed in actual environments, beyond the participants' ratings of realism.

The comparison between photographs and 360° video involved changes in temporal information, panoramic coverage, and viewpoint control. The observed differences therefore concern the presentation conditions as a whole and cannot be attributed to any one of these features. Spatial audio was used in all four scenarios, so the experiment assessed visual presentation under different background-sound conditions within a spatial-audio framework. In addition, the pairwise comparisons do not establish a formal interaction between visual presentation and background sound. Further factorial studies and analyses that account for repeated observations would help distinguish these contributions.

The comparative and correlation analyses focused on Node 3 as a common reference at the midpoint of the five-node sequence. This provided a consistent comparison location, but its position in the sequence does not establish that it represents the entire site or an optimal participant state. The spatial maps extend the descriptive presentation to all five nodes, but do not test differences between nodes. Further analyses accounting for repeated observations across nodes and scenarios are needed to assess spatial variation and the consistency of the findings. The questionnaire results also reflect subjective evaluations, and correlations among rating differences may partly involve shared response tendencies. They should therefore be interpreted as associations rather than evidence of causal relationships.

Finally, interaction relied on mouse-controlled viewing on a monitor, without head tracking. Changes in the virtual viewing direction and audio-listener orientation were consequently not coupled to the participants' physical head movements, which may affect perceived realism and comfort. Two out of the 38 recruited participants discontinued the experiment because of dizziness. As vestibular function and visual–vestibular weighting were not measured, these cases do not establish a specific mechanism of discomfort. Future comparisons of mouse-controlled and head-tracked interaction should include systematic assessments of both perceptual experience and simulator-sickness symptoms.

6. Conclusions

This study developed a monitor-based audiovisual simulation framework integrating recorded 360° video, spatial audio, and mouse-controlled viewing for built-environment perception evaluation. The framework coordinates visual exploration and spatial sound reproduction within a laboratory protocol, providing a methodological basis for comparing environmental experiences under defined presentation conditions.

The experiment examined differences between static photographs and 360° video under two background-sound conditions within this spatial-audio framework. At the analyzed node, the 360° video conditions received higher ratings for visual perception, perceived realism, and overall perception under both background-sound conditions. Differences in perceived richness and pleasantness were positively associated with differences in overall perception. These findings identify aspects of perceptual evaluation that vary with visual presentation and support the inclusion of audiovisual-characteristic ratings alongside scene-quality measures.

For built-environment research, the framework offers a basis for developing structured evaluations of public-space nodes and locations along pedestrian routes. Linking perceptual ratings with spatial context could help researchers and designers examine environmental experience in relation to landscape composition, spatial configuration, and sound sources. Such assessments can complement field measurements and observations when investigating the quality of existing places.

The present findings are specific to the university sample, campus setting, and selected analytical node. Further work should extend the analysis across locations and populations, assess alternative interaction methods, and compare laboratory ratings with in situ evaluations. These steps would clarify the framework's applicability to broader built-environment research and practice.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/buildings16193913/s1, Questionnaire S1: Blank version of Questionnaire 1 used in the study.

Author Contributions: Methodology, S.W.; software, Y.P. and S.W.; formal analysis, Y.P., H.Z. and J.X.; investigation, P.Y. and P.L.; data curation, P.Y. and P.L.; visualization, Y.P.; writing—original draft

preparation, Y.P. and H.Z.; writing—review and editing, Y.P., H.Z. and S.W.; supervision, S.W.; project administration, S.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Guangdong Provincial Social Science Planning Project (Grant No. GD26CYS37).

Informed Consent Statement: Written informed consent was obtained from all participants involved in the study.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request. The dataset is not publicly available due to privacy and ethical restrictions related to participant information.

Acknowledgments: The authors would like to thank all of the participants who contributed to the audiovisual evaluation experiment. During the preparation of this manuscript, the authors used GPT-5.6 Sol (OpenAI) to improve the English language and readability and to assist with translation. The authors reviewed and edited the AI-assisted content and take full responsibility for the published content.

Conflicts of Interest: Author Jian Xu was employed by the company Power China Huadong Engineering Corporation Limited. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Carles, J.L.; López Barrio, I.; de Lucio, J.V. Sound influence on landscape values. *Landsc. Urban Plan.* **1999**, *43*, 191–200. [CrossRef]
2. Gan, Y.; Luo, T.; Breitung, W.; Kang, J.; Zhang, T. Multi-sensory landscape assessment: The contribution of acoustic perception to landscape evaluation. *J. Acoust. Soc. Am.* **2014**, *136*, 3200–3210. [CrossRef] [PubMed]
3. Aletta, F.; Kang, J.; Axelsson, Ö. Soundscape descriptors and a conceptual framework for developing predictive soundscape models. *Landsc. Urban Plan.* **2016**, *149*, 65–74. [CrossRef]
4. Kang, J. Soundscape in city and built environment: Current developments and design potentials. *City Built Environ.* **2023**, *1*, 1. [CrossRef]
5. Liu, J.; Yu, S.; Wang, Y.; Zhang, J. Interactive effects between landscape characteristics and soundscape experience in urban parks. *Chin. Landsc. Archit.* **2017**, *33*, 86–90.
6. Cao, J.; Kang, J. The influence of companion factors on soundscape evaluations in urban public spaces. *Sustain. Cities Soc.* **2021**, *69*, 102860. [CrossRef]
7. Yang, X. Audio–Visual Interactions of Soundscape Elements in Urban Public Open Spaces. Ph.D. Dissertation, Harbin Institute of Technology, Harbin, China, 2023. [CrossRef]
8. Kang, J. *Urban Sound Environment*; CRC Press: London, UK, 2006.
9. Zhao, W.; Li, H.; Zhu, X.; Ge, T. Effect of birdsong soundscape on perceived restorativeness in an urban park. *Int. J. Environ. Res. Public Health* **2020**, *17*, 5659. [CrossRef] [PubMed]
10. Lee, P.J.; Park, S.H.; Jung, T.; Swenson, A. Effects of exposure to rural soundscape on psychological restoration. In *European Congress and Exposition on Noise Control Engineering (Euronoise)*; EAA: Crete, Greece, 2018.
11. Ren, X.; Li, Q.; Yuan, M.; Shao, S. How visible street greenery moderates traffic noise to improve acoustic comfort in pedestrian environments. *Landsc. Urban Plan.* **2023**, *238*, 104839. [CrossRef]
12. Yang, W.; Kang, J. Soundscape and sound preferences in urban squares: A case study in Sheffield. *J. Urban Des.* **2005**, *10*, 61–80. [CrossRef]
13. Yang, W.; Kang, J. Acoustic comfort evaluation in urban open public spaces. *Appl. Acoust.* **2005**, *66*, 211–229. [CrossRef]
14. Lian, Y.; Ou, D.; Pan, S. A comparative study of soundscape evaluation between laboratory simulation and field surveys. *Build. Sci.* **2021**, *37*, 139–144. [CrossRef]
15. Hermida Cadena, L.F.; Lobo Soares, A.C.; Pavón, I.; Coelho, L.B. Assessing soundscape: Comparison between in situ and laboratory methodologies. *Noise Mapp.* **2017**, *4*, 57–66. [CrossRef]
16. Preis, A.; Kociński, J.; Hafke-Dys, H.; Wrzosek, M. Audio-visual interactions in environment assessment. *Sci. Total Environ.* **2015**, *523*, 191–200. [CrossRef] [PubMed]
17. Lindquist, M.; Lange, E.; Kang, J. From 3D landscape visualization to environmental simulation: The contribution of sound to the perception of virtual environments. *Landsc. Urban Plan.* **2016**, *148*, 216–231. [CrossRef]

18. Nordahl, R. Evaluating environmental sounds from a presence perspective for virtual reality applications. *EURASIP J. Audio Speech Music Process.* **2010**, *2010*, 426937. [CrossRef]
19. Jeon, J.Y.; Jo, H.I. Effects of audio-visual interactions on soundscape and landscape perception and their influence on satisfaction with the urban environment. *Build. Environ.* **2020**, *169*, 106544. [CrossRef]
20. Jeon, J.Y.; Jo, H.I.; Santika, B.B.; Lee, H. Crossed effects of audio-visual environment on indoor soundscape perception for pleasant open-plan office environments. *Build. Environ.* **2022**, *207*, 108512. [CrossRef]
21. Potter, T.; Cvetković, Z.; De Sena, E. On the relative importance of visual and spatial audio rendering on VR immersion. *Front. Signal Process.* **2022**, *2*, 904866. [CrossRef]
22. Cao, J.; Kang, J. A perceptual structure of soundscapes in urban public spaces using semantic coding based on the grounded theory. *Int. J. Environ. Res. Public Health* **2023**, *20*, 2932. [CrossRef] [PubMed]
23. Georgiou, F.; Kawai, C.; Schäffer, B.; Pieren, R. Replicating outdoor environments using VR and ambisonics: A methodology for accurate audio-visual recording, processing and reproduction. *Virtual Real.* **2024**, *28*, 111. [CrossRef] [PubMed]
24. Seamon, D. Merleau-Ponty, Environmental Embodiment, and Place: Implications for Architecture and Placemaking. In *Embodied Awareness of Space*; Springer Nature: Singapore, 2024; pp. 97–122. [CrossRef]
25. Drever, J.L. Primacy of the Ear'—But Whose Ear? The case for auraldiversity in sonic arts practice and discourse. *Organised Sound* **2019**, *24*, 85–95. [CrossRef]
26. Böhme, G. Atmosphere as a fundamental concept of a new aesthetics. In *The Aesthetics of Atmospheres*; Routledge: London, UK, 2016; pp. 23–36. [CrossRef]
27. Böhme, G. Atmosphere as the subject matter of architecture. In *The Aesthetics of Atmospheres*; Routledge: London, UK, 2016; pp. 147–152. [CrossRef]
28. Schulte-Fortkamp, B.; Jordan, P. When soundscape meets architecture. *Noise Mapp.* **2016**, *3*, 216–321. [CrossRef]
29. Lu, Y.; Hasegawa, Y.; Tan, J.K.A.; Lau, S.-K. Effect of audio-visual interaction on soundscape in the urban residential context: A virtual reality experiment. *Appl. Acoust.* **2022**, *192*, 108717. [CrossRef]
30. Gök Tokgöz, Ö.; Engel, M.S.; Othmani, C.; Altinsoy, M.E. Virtual Reality Application in Evaluating the Soundscape in Urban Environment: A Systematic Review. *Acoustics* **2025**, *7*, 68. [CrossRef]
31. Latini, A.; Torresin, S.; Oberman, T.; Di Giuseppe, E.; Aletta, F.; Kang, J.; D'Orazio, M. Virtual reality application to explore indoor soundscape and physiological responses to audio-visual biophilic design interventions. *J. Build. Eng.* **2024**, *87*, 108947. [CrossRef]
32. Jeon, J.Y.; Santika, B.B.; Lee, H.; Yoon, H.; Ahn, D.; Kim, J.S. Music and soundscape effects on young adults' psychophysiological restoration in virtual reality: A longitudinal study. *Appl. Acoust.* **2026**, *248*, 111282. [CrossRef]
33. Oberman, T.; Mitchell, A.; Aletta, F.; Kang, J. How does visual context of urban open spaces drive soundscape pleasantness and eventfulness: A laboratory experiment using a higher order Ambisonics speaker system and immersive virtual reality. *INTER-NOISE NOISE-CON Congr. Conf. Proc.* **2024**, *270*, 6350–6356. [CrossRef]
34. Cerviño, J.; Lombera, E.N.; Piceda, L.B.; Eguia, M.C.; Vergara, R.O. The influence of visual context on auditory distance perception. *Appl. Acoust.* **2026**, *250*, 111334. [CrossRef]
35. Chen, X.; Kang, J. Effects of natural sound tempo and social conditions on social interaction: An interactive virtual experiment. *J. Environ. Psychol.* **2025**, *104*, 102630. [CrossRef]
36. Jiang, J.; Meng, Q.; Yang, D.; Li, M.; Kang, J. Simulating spatial experience in exhibition spaces with musical soundscapes using virtual reality. *Build. Simul.* **2025**, *18*, 1497–1514. [CrossRef]
37. Mi, T.-W.; Yang, M.-T. Comparison of tracking techniques on 360-degree videos. *Appl. Sci.* **2019**, *9*, 3336. [CrossRef]
38. Lee, J.-K.; Lee, S.; Kim, Y.; Kim, S.; Hong, S.-W. Augmented virtual reality and 360 spatial visualization for supporting user-engaged design. *J. Comput. Des. Eng.* **2023**, *10*, 1047–1059. [CrossRef]
39. Zhao, X.; Lin, G. Research on the perception evaluation of urban green spaces using panoramic images and deep learning: A case study of Zhujiang Park in Guangzhou. *Landsc. Archit. Front.* **2024**, *12*, 7. [CrossRef]
40. Wightman, F.L.; Kistler, D.J. Headphone simulation of free-field listening. I: Stimulus synthesis. *J. Acoust. Soc. Am.* **1989**, *85*, 858–867. [CrossRef] [PubMed]
41. Wightman, F.L.; Kistler, D.J. Headphone simulation of free-field listening. II: Psychophysical validation. *J. Acoust. Soc. Am.* **1989**, *85*, 868–878. [CrossRef] [PubMed]
42. Zotter, F.; Frank, M. *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*; Springer: Cham, Switzerland, 2019. [CrossRef]
43. Funkhouser, T.; Min, P.; Carlbom, I. Real-time acoustic modeling for distributed virtual environments. In *Proceedings of SIGGRAPH '99*; ACM Press: Los Angeles, CA, USA, 1999; pp. 365–374. [CrossRef]
44. Zahorik, P. Perceptually relevant parameters for virtual listening simulation of small room acoustics. *J. Acoust. Soc. Am.* **2009**, *126*, 776–791. [CrossRef] [PubMed]
45. Hirway, A.; Qiao, Y.; Murray, N. Spatial audio in 360° videos: Does it influence visual attention? In *Proceedings of MMSys '22*; ACM: Athlone, Ireland, 2022; pp. 39–51. [CrossRef]

46. Drever, J.L.; Hugill, A. (Eds.) *Aural Diversity*; Routledge: London, UK, 2022. [CrossRef]
47. Kang, J.; Fang, X.; Aletta, F.; Mitchell, A.; Oberman, T.; Cerwén, G.; Chen, X.; Xiao, J.; Mittal, P. Soundscape Practice and Interventions: A Framework. In Proceedings of the 11th Convention of the European Acoustics Association Forum Acusticum/EuroNoise 2025, Málaga, Spain, 23–26 June 2015; pp. 4699–4705. [CrossRef]
48. Lavia, L.; Dixon, M.; Witchel, H.J.; Goldsmith, M. Applied Soundscape Practices. In *Soundscape and the Built Environment*; Kang, J., Schulte-Fortkamp, B., Eds.; CRC Press: Boca Raton, FL, USA, 2015; pp. 243–301. [CrossRef]
49. Xiao, J.; Lavia, L.; Kang, J. Towards an agile participatory urban soundscape planning framework. *J. Environ. Plan. Manag.* **2018**, *61*, 677–698. [CrossRef]
50. Bruschi, V.; Grossi, L.; Dourou, N.A.; Quattrini, A.; Vancheri, A.; Leidi, T.; Cecchi, S. A review on head-related transfer function generation for spatial audio. *Appl. Sci.* **2024**, *14*, 11242. [CrossRef]
51. Huang, Q.; Li, Z. Examining immersive audio's role in flow and gaze behavior in pseudo-VR environments. *Front. Virtual Real.* **2025**, *6*, 1691405. [CrossRef]
52. Liu, J.; Wang, Y.; Zimmer, C.; Kang, J.; Yu, T. Factors associated with soundscape experiences in urban green spaces: A case study in Rostock, Germany. *Urban For. Urban Green.* **2019**, *37*, 135–146. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.