Scientific and Social Research, 2026, Volume 8, Issue 8
http://ojs.bbwpublisher.com/index.php/ssr
Online ISSN: 2981-9946
Print ISSN: 2661-4332

Generative Reconstruction of Intangible Cultural Heritage Soundscapes from the Perspective of Spatial Computing

Jiasheng Jin*

Sichuan Film and Television University, Chengdu 610036, Sichuan, China

Author to whom correspondence should be addressed.

Abstract: In response to the de-spatialization of intangible cultural heritage (ICH) sounds and the dissolution of the audience's embodied perception during the digital dissemination of traditional media, this paper takes Gyalrong Tibetan polyphonic folk songs as an example to explore a generative reconstruction pathway for ICH soundscapes from the perspective of spatial computing. Through fieldwork, auditory, visual, and spatial data were collected and aligned using BeiDou Companion, Blender, and QGIS to construct a multimodal dataset. AIGC technology was introduced: after sound source separation, multimodal features were fused to perform generative modeling of a first-order Ambisonics sound field, and in combination with an acoustic propagation model, spatial reconstruction and binaural rendering were accomplished. Controlled experiments grounded in embodied theory demonstrate that the spatialized soundscape significantly outperforms traditional stereo in terms of immersion, cultural resonance, and physiological arousal. Consequently, this paper proposes an "embodied empathy" mechanism, offering a new research paradigm for the living transmission of ICH in the AI era.

Keywords: ICH soundscapes; Polyphonic folk songs; Embodied communication; Gyalrong Tibetan; Spatial computing

Online publication: August 31, 2026

1. Introduction

1.1. Research background

Within the multi-ethnic cultural landscape of southwest China, the polyphonic folk songs of the Gyalrong Tibetan people stand out among numerous intangible cultural heritage (ICH) elements by virtue of their distinctive melodies and spatialized performance practice—a rare and highly integrated form of sonic culture. Such folk songs are typically born in everyday settings and ritual arenas, such as cultivated fields or around campfires; their multiple vocal parts are not simply superimposed but form a dynamic sound-field structure arising from specific spatial relationships [1]. In short, Gyalrong Tibetan polyphonic folk songs are rooted in

the synergistic relationship between space and the participants (the singers). Regrettably, in the current digital dissemination of ICH, their sound relies mainly on two-channel or even monaural recording, with video imagery taking a dominant role. While these "two-dimensional media" offer clear advantages in information storage and transmission efficiency, they exhibit a dimensional deficiency in recording and digitally reproducing the spatial dimension of sound. For polyphonic folk songs, using two-channel stereo to record and disseminate them compresses the spatial relationships among the multiple vocal parts, weakens the interaction between sound sources and the environment, and renders the original atmospheric quality almost impossible to convey effectively. At the present stage, the communication situation is that highly spatial ICH soundscapes are being "de-spatialized" in the process of dissemination, and the cultural connotations embedded in polyphonic folk songs are simultaneously diluted [2].

With the development of artificial intelligence (AI) and spatial computing technologies in recent years, the problem of ICH soundscapes being "de-spatialized" during dissemination is gradually being resolved, and the ways in which sound is digitally processed and communicated have undergone a fundamental shift [3]. Emerging technologies, represented by three-dimensional sound-field modeling, binaural rendering, and intelligent generative models, make it possible for ICH soundscapes to move from "documentary archiving" toward "spatial reconstruction." In this context, how to reconstruct the spatial attributes of Gyalrong Tibetan polyphonic folk songs by means of these new technological tools has become a critical issue demanding urgent attention in current ICH digitization research.

1.2. Origin of the problem

Conventional recording technology centers on the capture and reproduction of sound waveforms, paying primary attention to parameters such as frequency, amplitude, and pitch, while neglecting the position and movement of sound in three-dimensional space and its coupling relationship with the environment. In this process, the perceptual system constituted by sound source location, spatial structure, and environmental sound—namely the auditory sphere—is reduced to a linear audio signal [4]. For polyphonic folk songs, the "dissolution" effect brought about by this medium is manifested in three main aspects. First, the spatial distribution relationships among the multiple vocal parts are compressed into a single sound image, causing a sound system that originally possessed directional differences and a hierarchical structure to become flattened. Second, the interactive relationship between the sound and the environment is weakened, making it difficult to present the original spatial texture. Third, the listener transforms from an "embodied participant" into an "external observer," and the mode of perception shifts from immersive experience to informational reception.

Under current technological conditions, therefore, it remains difficult for ICH sound scenes to evoke a sense of being physically present; their spatial information is stripped away, and the audience's bodily perception is severed. This paper seeks to explore a new technological pathway capable of reconstructing this "auditory sphere" of ICH in the "digital world" and restoring its complete cultural expression.

1.3. Spatial computing

In the face of the deficiencies of traditional technical means in spatial expression, spatial computing technologies, which have risen to prominence in recent years, provide a new technical foundation for the spatial digital reconstruction of sound. Different from the former paradigm of linear recording and playback,

spatial computing emphasizes the three-dimensional construction and dynamic simulation of the physical world; its core lies in using computation to "build," with data, the way a recorded object exists in three-dimensional space.

In the present study, the application of spatial computing involves the following principal aspects. First, using three-dimensional sound-field modeling techniques to transform the spatial structure of polyphonic folk songs into adjustable and computable parameters. Second, through sound source localization and coordinate mapping, accurately "placing" the different vocal parts into a computable spatial framework, restoring them to their proper positions. Third, utilizing acoustic propagation models to simulate how sound is reflected, diffracted, and generates reverberation within a specific environment, thereby approximating the original soundscape. Concurrently, the development of AI technology accelerates the realization of this pathway. Unlike traditional sound signal processing methods, large language models can infer and complete missing information by learning the distributional relationships within large volumes of data. This means that in situations where the original sound information is incomplete, AI possesses the capacity to reconstruct ICH soundscapes by learning from and modeling multimodal data. As a result, the media conversion of ICH sounds will undergo a qualitative leap, moving from recording aimed at "archiving" in the past to generation centered on "reconstruction" now, and from linear repetition aiming at "reproduction" to multidimensional modeling with "experience" at its core.

1.4. Research methods

Based on the above problems and the technical means adopted, this paper employs an interdisciplinary research methodology to construct a canonical research pathway from "field investigation" to "reconstruction and empirical verification" [5]. First, at the level of data collection, the research uses participatory video observation to carry out in-depth recording of the actual performance contexts of Gyalrong Tibetan polyphonic folk songs [6]. By deeply engaging in field activities and establishing long-term collaborative relationships with ICH inheritors, the author collects audio, video, and spatial information of the polyphonic folk songs in authentic environments, thereby ensuring the authenticity and completeness of the data. Subsequently, in the data processing stage, supported by multimodal data, a structured soundscape dataset is constructed [7]. Through technical means such as sound source separation, spatial annotation, and environment modeling, the raw collected data are transformed into a computable soundscape representation, providing a foundation for the subsequent generative model. Following this, Artificial Intelligence Generated Content (AIGC) is introduced to model and reconstruct the spatial sound field of the polyphonic folk songs; by combining acoustic propagation principles with neural network models, the spatialized generation of the original soundscape is realized, and within a binaural hearing framework, an immersive audio result is output [8].

Finally, at the stage of communication validation, the study conducts perceptual experiments based on embodied cognition theory, performing comparative analyses of the differences between spatial audio and traditional stereo audio in dimensions such as sense of space, immersion, and emotional resonance, thereby verifying the effectiveness of generative reconstruction in the dissemination of polyphonic folk songs. Through the adoption of these methods, this paper strives to address, both theoretically and practically, the two questions of "how to reconstruct" and "how to perceive" ICH soundscapes, offering a new research paradigm for the living transmission of intangible cultural heritage in the AI era [9].

1.5. Research hypothesis

On the foundation of the aforementioned soundscape ecology and embodied cognition theories, it is not difficult to see that the core problem in the communication of ICH soundscapes can be attributed to the loss of spatial information and the disruption of the perceptual mechanism. The study posits that AI technology, by learning from and integrating multiple sensory modalities, is able to construct a virtual auditory environment in the digital world, thereby supplementing the sound-field information missing from the original recording. When listeners are immersed in this virtual space and experience the sound, their bodily sensations are re-engaged, enabling them to perceive a sense of presence akin to actually being there. In this way, emotional resonance is more readily evoked, cultural understanding is deepened, and the effectiveness of ICH communication is thereby enhanced. The significance of this idea lies in the fact that introducing AI into ICH research makes it no longer merely a cold technical tool, but rather something more like a bridging intermediary connecting sound structure with bodily experience. The key is not to replicate sound exactly as it was, but to reconstruct the living relationship among sound, space, and the body. Through this process, the ICH soundscape is transformed from "a documented object" into "a personally experienceable field," thus acquiring an entirely new vitality in the digital world.

2. Theoretical foundations

2.1. The spatial structure of polyphonic folk songs

In the course of sound theory development, the soundscape is understood as a holistic auditory system co-constituted by the natural environment, human activities, and cultural behaviors. Compared with traditional physical acoustics, which treats sound as a physical signal, soundscape ecology places greater emphasis on the dynamic relationship between sound and its surrounding environment, as well as the listener's perceptual position and mode of participation within this relational process. In this way, sound is not merely an object to be recorded, but a cultural form that permeates specific geographical and social contexts.

Gyalrong Tibetan polyphonic folk songs are formed through group singing. The spatial distribution of the singers is not random, but closely bound up with labor scenes, ritual forms, and social relations. The positional differences among the singers give rise, during the propagation of sound, to a distinctly stratified auditory structure. Such polyphonic structures unfold within specific spaces, thereby constituting a perceivable sound-field order (Figure 1). Looking further, the soundscape of Gyalrong Tibetan polyphonic folk songs also forms a deep coupling relationship with the geographical environment. The echo effects brought about by mountainous terrain, the absorption of high-frequency sounds by forested environments, and the influence of village architecture on sound-wave reflection paths—all these factors imperceptibly participate in the process of musical generation. Thus, sound not only propagates in space but is also shaped by space. In other words, the artistic form of the folk songs does not exist independently of the environment, but is continuously generated through the interaction among sound, geography, and the body.

Figure 1. Natural soundscape and spatial interaction scene of Gyalrong Tibetan polyphonic folk songs (Photograph location: Qiaoqi Tibetan Township, Baoxing County, Ya'an City, Sichuan Province; Photo credit: taken by the author)

In this regard, the recording of polyphonic folk songs through traditional media formats (two-channel stereo audio) is, in effect, a "stripping away" of their soundscape. When the spatial dimension is compressed, the original sound-field relationships collapse, and the holistic perception of the music is dissociated into discrete auditory fragments. Therefore, from the perspective of soundscape ecology, the spatial reconstruction of Gyalrong Tibetan polyphonic folk songs constitutes both a restoration of their artistic structure and a reconnection with their cultural context.

2.2. Embodied communication theory

Within communication studies, embodied cognition provides a new theoretical framework for understanding perception and meaning-making. This theory posits that human cognitive activity does not occur solely within the abstract brain, but depends on the continuous interaction between the body and the environment. Perception is not merely a process of information reception, but the very means by which the body locates itself in space and participates in the world. Introducing this theory into the field of sound communication, it is readily apparent that auditory experience possesses strong spatial attributes. By detecting interaural time differences and interaural level differences, an individual can determine the location, distance, and trajectory of a sound source, thereby forming a complete spatial perception. Consequently, the communicative effect of sound depends not only on its acoustic features, but more crucially on its capacity to activate the listener's bodily experience.

In the context created by Gyalrong Tibetan polyphonic folk songs, the listener is not merely an external observer but can also encompass the singers situated within the sound field. Sound arrives from various directions, producing an enveloping or flowing effect within the space, thereby generating a sense of position and participation in the individual's auditory experience. Such an experience is difficult to reproduce through traditional stereo recording, the reason being that stereo recording lacks the capacity to finely articulate the spatial relationships among sound sources and cannot simulate the dynamic propagation of sound within an

environment.

Embodied communication theory further points out that a very close connection exists between perception and emotion. When an individual achieves a sense of spatial presence auditorily, their emotional resonance typically becomes more intense. In other words, the communication of intangible cultural heritage sounds requires not only the preservation of their melody and rhythm, but also the reconstruction of their "body-perceptible" spatial structure. Only by achieving reproduction on an embodied level can the transmission of cultural meaning reach a depth of empathic resonance.

3. Empirical basis

3.1. Field investigation

As the principal field researcher and data owner for this project, the author arrived in Ya'an, Sichuan, on April 10, 2026, and conducted investigations in Qiaoqi Tibetan Township through April 16. During this period, participant observation methods were employed to gather data in depth from the field settings (Figure 2). On April 10, a cultural survey was carried out in the Qiaoqi Tibetan village, with the entire process recorded using a Sony A7M4 camera and a Zoom F8n professional audio recorder. On April 11, at the home of inheritor Ms. Wang Lianjiang, material was recorded of several inheritors singing folk songs together. On April 12, in the small courtyard of her home, a scene of singing spontaneously accompanying daily labor was documented. On April 13, an indoor gathering and Guozhuang dance at her home were recorded. On April 14, a polyphonic folk song performance performed by fourteen people together was captured on a hillside. On April 15 and 16, in-depth interviews were further conducted at Ms. Wang Lianjiang's home, and additional recording of the relevant cultural context was undertaken through communal living and close observation.

For the recording of the outdoor fourteen-person polyphonic folk song performance on April 14, this study used a multi-track recorder with XLR inputs (Zoom F8n), equipped with omnidirectional microphones (six RODE NTG5), deploying a research-grade microphone array to cover the principal sound source directions (Figure 2).

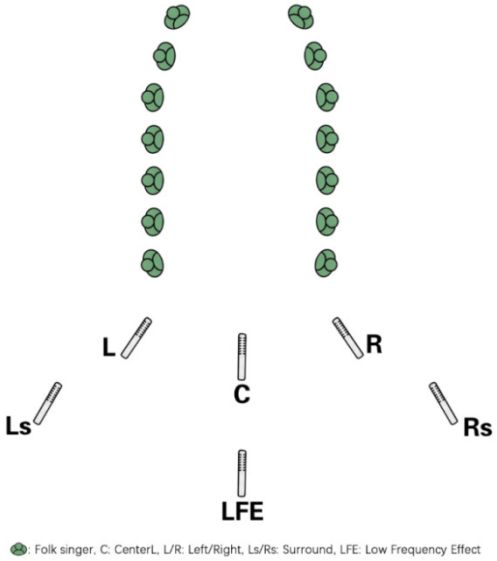

Figure 2. Schematic diagram of the spatial distribution of the microphone array for recording the fourteen-person polyphonic folk songs

3.2. Multimodal dataset

Based on the data collection implemented above, a dataset containing three layers of information was organized: audio, visual, and spatial. The audio layer primarily contains the multi-track recordings of the folk song performances. Each original file is in WAV format, with a 48 kHz sampling rate, 24-bit depth, and 6 channels. The visual layer consists of high-definition video clips recorded synchronously with the audio; each video file is in MOV format with a resolution of 1920×1080, a frame rate of 60 fps, and includes corresponding audio tracks or index annotations.

The spatial layer mainly records the positional relationships on site. First, spatial data were collected using BeiDou Companion software to measure the relative three-dimensional coordinates between the singers and the center points of the Tibetan village and the signboard, establishing a local Cartesian coordinate system. Subsequently, the recorded video clips were imported into Blender software, which automatically identified feature points in the footage and solved for the three-dimensional motion trajectory and spatial position of the moving camera. Then, the previously measured three-dimensional spatial coordinates were imported into an open-source Geographic Information System (QGIS). Using the local coordinate system established based on these coordinates as a calibration reference, the visually solved camera motion trajectories were imported. By comparing them against the static coordinates physically measured by BeiDou Companion, any drift errors occurring during the lens tracking process were corrected to complete the spatial alignment and calibration. In this way, the accurate spatial position information of the singers was obtained and stored as a JSON file for subsequent program calls and 3D scene importing.

3.3. Data and algorithms

After data collection, all audio data were preprocessed to facilitate later computational processing. This primarily includes four parts: sound source separation, environmental noise reduction, acoustic feature extraction, and spatial annotation. For sound source separation, a pre-trained deep learning model—the Spleeter tool released by Deezer—was used to split the mixed recordings into vocal and accompaniment parts. In the noise reduction stage, spectral subtraction or an open-source algorithm based on Wiener filtering was adopted to attenuate background noise to a low audibility level. All audio was analyzed frame-by-frame in the time-frequency domain using the Short-Time Fourier Transform (STFT) with a window length of 1024 points (≈21.3 ms) and a frame shift of 512 points (50% overlap) to extract power spectrograms. The extracted acoustic features include Mel-Frequency Cepstral Coefficients (MFCC, 13 dimensions), spectral centroid, spectral roll-off, and short-time energy, among others, for subsequent perceptual analysis and machine learning. Where auditory spatial perception needed to be considered, Head-Related Transfer Functions (HRTFs) were collected or simulated to prepare binaural rendering data.

In the process of spatial information annotation, all sound source positions were calibrated in meters within the local coordinate system. BeiDou Companion was used to measure the distances between the fixed singers and reference points, and then angular information was converted into coordinates based on the positions of the sound sources in the video. If a sound source moved, its trajectory was dynamically calculated by combining video frames and multi-channel time delay differences. The pseudocode is shown in Table 1.

Table 1. Algorithm 1: Spatial coordinate calibration and dynamic source tracking

Algorithm 1: Spatial Coordinate Calibration & Dynamic Source Tracking

REQUIRE: multi-channel_audio, video_frames, camera_intrinsics, mic_array_geometry
ENSURE: spatial_trajectory (x, y, z, t)

```
// Global Reference: Define origin (0,0,0) at the center of the microphone array
for each timestamp t do
    // Static Calibration
    distance ← laser_rangefinder.read()
    angle ← detect_object(video_frames[t], object_id) // Using computer vision
    spatial_pos ← polar_to_cartesian(angle, distance)

    // Dynamic Tracking (Sensor Fusion)
    // TDOA calculates relative delay between N microphones
    audio_direction ← compute_tdoa(multi_channel_audio[t])
    // Apply Extended Kalman Filter to combine video and audio inputs
    trajectory[t] ← ekf_update(spatial_pos, audio_direction, t)
    save_spatial_label(object_id, trajectory[t], t)
end for
```

In the process of data quality control, each recording was strictly checked for clipping (i.e., ensuring no overload), and signal synchronization checks were performed to guarantee time alignment across all channels. Information such as the recordist, date, and equipment was documented for each file to ensure data traceability. The data processing pipeline was recorded through script automation, with a list of key steps and sample code provided in the documentation to enable experimental reproducibility. Technical risks mainly included power depletion of recording devices during long sessions and sudden weather changes affecting recording quality. To address these, spare batteries and waterproof covers were prepared, and ambient noise levels were constantly monitored during outdoor shooting to avoid the risk of data loss or quality degradation. In terms of ethics, all participants fully understood the purpose of the recording and cooperated actively. Audio and video content containing sensitive information was de-identified to ensure respect for the inheritors' privacy and authenticity.

4. Technical mechanisms

4.1. Generative sound field modeling

At present, four mainstream categories of generative models predominate. The first is the diffusion model, whose basic idea is to "add noise to the data and then learn to remove it step by step." This approach has proven highly effective in restoring complex high-dimensional sound field distributions and is frequently used in audio synthesis, especially spatial audio synthesis. The second is the Variational Autoencoder (VAE), which, through an "encode-then-decode" architecture, can automatically learn the most critical latent variable features from the sound field. In the absence of explicit labels, this method is well suited for extracting spatial information from sound. The third category is conditional Transformers, which can accept textual or visual prompts as conditions, thereby imposing semantic or location-specific constraints on the generated audio, so that the resulting sound better matches the intended scene. The fourth is physics-guided neural networks.

To effectively guide the model in understanding the generated sound field, this study provides three types of multimodal information as inputs: multi-channel waveforms, spatial cues (including binaural hearing cues such as ITD and ILD), and visual features. The sound field is represented using the first-order Ambisonics (FOA) format, which records complete spatial information via four channels. The model fuses

these multimodal inputs to form rich generative conditions, with the goal of synthesizing a high-fidelity spatial sound field. To ensure generation quality while considering both auditory perception and physical metrics, this study designed a composite loss function. Its core component is the Multi-Resolution Short-Time Fourier Transform loss (MS-STFT Loss). This loss compares the spectral differences between the generated and real audio across multiple sets of time-frequency resolutions, thereby ensuring similarity in pitch, timbre, and other spectral characteristics. Since the sound field is represented by four FOA channels (W, X, Y, and Z), the MS-STFT loss is calculated separately for each channel and then averaged.

4.2. Spatial computing

Accurately mapping live-recorded sound to real spatial coordinates and effectively integrating it into an acoustic propagation model is a crucial step in reconstructing an auditory field with a genuine sense of space. This study constructed a local Cartesian coordinate system, taking Qiaoqi Lake in Qiaoqi Tibetan Township as the origin. After obtaining the geographic coordinates of the singers and on-site reference objects via BeiDou Companion, these geographic coordinates were calibrated using multi-point triangulation and then transformed into a unified planar coordinate system. For dynamic sound sources during the performance, their coordinates could be acquired through video analysis and combined with multi-view 3D reconstruction to deduce the continuous motion trajectories of the singers. The overall concept of this coordinate mapping draws on the FoleySpace algorithm, whose core principle is to map two-dimensional positions in the image into a three-dimensional sound field coordinate system.

Next, the acoustic propagation model is introduced. To make the sound more realistic, it is necessary to consider various propagation effects that occur as sound travels through space, such as direct sound, first-order reflections, diffraction, and overall reverberation. Generally, physical simulation methods can be employed: for instance, ray tracing can simulate direct sound and reflections, the image source method can calculate specific reflection paths, or a reverberation time model (i.e., the Sabine formula) can estimate the overall reverberation characteristics of the space. Within a generative framework, these physical factors can be integrated into the model in different ways. For example, room geometry parameters (such as dimensions and reflection coefficients) can be fed as conditional inputs, allowing the network itself to learn how sound should be reflected and attenuated. Alternatively, Physics-Informed Neural Networks (PINNs) can be used, incorporating the wave equation or energy constraints into the loss function to ensure the generated sound field conforms to physical laws.

Finally, by combining the discrete sound source coordinates obtained through the preceding series of operations with this acoustic propagation model, a complete digital sound field can be realized. The audio ultimately generated naturally contains information about the direction and distance of the sound sources, laying the foundation for subsequent binaural rendering and immersive listening.

4.3. Algorithm-assisted narrative reconstruction

Building on the previously generated model, and to address the issue of missing information, this study posits that audio from adjacent time segments can be used for gap-filling through linear interpolation or neural network-based vocoders. If some channel signals are lost, the sound priors provided by sound source separation can be leveraged to recover and reconstruct unrecorded vocal parts. The model can also perform guided audio inpainting. For example, by inputting relevant syllables or semantic keywords, a conditional

generative model can be driven to synthesize corresponding audio segments, achieving semantic-based local completion.

In research on multimodal alignment, this study utilizes audio-visual correspondence models to ensure synchronization between the generated audio and the lip movements and actions of figures in the video. Specifically, a Transformer architecture with time encoding can be adopted, taking the action features or labels of the sound source as conditional input, thereby guiding the generated audio sequence to remain temporally consistent with the visual stream. Furthermore, the study considers that introducing narrative-driven editorial control can make the auditory presentation of ICH stories more vivid and emotionally compelling. For instance, at key lyrics or emotional climaxes, the spatial effect or reverberation of the sound field can be enhanced, while during non-climactic passages, background sounds congruent with the environment can be timely incorporated, thereby reinforcing the narrative rhythm and emotional tension within the spatialized expression and elevating the overall immersive experience. The pseudocode is shown in Table 2.

Table 2. Algorithm 2: Audio frame processing and restoration

Algorithm 2: Audio Frame Processing & Restoration
REQUIRE: raw_audio_stream, spatial_metadata, event_timestamps ENSURE: restored_and_enhanced_audio

// Initialize buffer for Overlap-Add (OLA)
for each time frame t with overlap do
  if is_corrupted(t) then
  context ← get_surrounding_audio(t)
  interpolated ← generative_model.predict(context)
  frame_data ← apply_hann_window(interpolated)
  else
  frame_data ← get_audio(t)
  end if
 	
  for each source in active_sources do
  // Apply beamforming with coefficient smoothing
  frame_data ← spatial_beamformer(frame_data, source_pos[t])
  end for
 	
  if is_key_event(t) then
  frame_data ← apply_dynamic_range_compression(frame_data)
  end if
 	
  output_buffer ← overlap_add(output_buffer, frame_data)
end for

5. Communication experiment

5.1. Experimental design

Through a controlled experiment, this study aims to verify whether the generative soundscape reconstruction technology based on spatial computing can yield a better experience effect in ICH communication compared with traditional audio, and simultaneously to test the applicability of relevant cognitive models. The following testable hypotheses are proposed: (1) Enhanced spatial presence, i.e., the group experiencing the generative spatialized soundscape should score significantly higher on the sense of presence than the group

listening to traditional audio; (2) Improved immersion and cultural resonance, i.e., the experimental group should score significantly higher than the control group on subjective evaluations such as immersion and emotional resonance; (3) Differences in physiological arousal, i.e., the physiological indicators such as heart rate and skin conductance of participants listening to spatial audio should reflect a higher level of excitation, indicating stronger bodily engagement; (4) Differences in behavioral engagement, i.e., participants in the spatial audio group may exhibit longer fixation duration and richer head movements during the listening process, reflecting a higher level of bodily interaction.

Regarding participants, this study recruited adults aged 18 to 35, ensuring normal hearing and no relevant medical history, with a roughly balanced gender ratio. Through channels such as campus recruitment, 60 participants were eventually randomly invited. The sample size was determined by a prior power analysis, sufficient to test the above hypotheses.

The experiment employed a two-group controlled design. The experimental group listened to spatialized polyphonic audio processed through generative technology, while the control group listened to stereo audio mixed conventionally. Participants were randomly assigned to the two groups, with balance ensured for characteristics such as gender and age. The audio materials used in the experiment were all derived from polyphonic folk songs collected in the field in Qiaoqi, with a sampling rate of 48 kHz and a bit depth of 24 bits. The spatialized audio was generated by applying room impulse responses to each vocal part and combining binaural rendering techniques to produce two-channel stereo; the control group's audio was a conventional stereo mix. All audio was uniformly normalized before playback to ensure consistent loudness. The experiment was conducted in a soundproof laboratory with soft lighting and a suitable temperature to avoid interference. The same model of high-fidelity headphones and a professional external sound card were used as playback equipment, with the volume uniformly set at an average of approximately 75 dB SPL, and a quiet ambient background was maintained to ensure identical listening conditions for both groups.

5.2. Data analysis

During the data collection process, after each participant finished listening to the audio, this study simultaneously collected subjective questionnaire feedback as well as multi-channel physiological data and behavioral data. The questionnaire was filled out via a computer interface, while the physiological and behavioral signals were uniformly acquired and recorded by the experimental computer. To temporally align all data, a digital trigger signal was simultaneously sent to the physiological acquisition system and the eye tracker at the moment of audio playback, thereby placing precise timestamps in each data stream. All recording devices were also synchronized via a common clock or the Network Time Protocol to reduce errors caused by clock drift.

In terms of statistical testing methods, this study selected appropriate analytical methods based on the different groups and variable types. For instance, for subjective rating data, repeated measures ANOVA was used if the same participant listened to different audio as a within-group comparison; independent samples t-test or one-way ANOVA was used for between-group comparisons. For physiological and behavioral data, mixed-effects models could be adopted to better incorporate individual differences among participants as random effects. When reporting statistical results, besides examining the P-value, the corresponding effect size was also calculated. For example, Cohen's d was used for t-tests, and partial eta squared for ANOVA, in order to more comprehensively judge the practical significance of the differences.

The quality of the questionnaire scales was also examined. First, the KMO measure and Bartlett's test of sphericity were used to determine whether the data were suitable for factor analysis. Then, methods such as principal component analysis were employed to extract factors, generally retaining factors with eigenvalues greater than 1 and aiming for a cumulative explained variance exceeding 70%. Next, Cronbach's α coefficient was used to assess the internal consistency of the scale, typically requiring $\alpha > 0.7$. If further confirmation of the scale structure was needed, confirmatory factor analysis could also be performed. Finally, all analyses set the alpha level at 0.05, and multiple comparison corrections were applied when necessary, using R software (v4.5.2) for statistical analysis.

5.3. Results and verification

Before the formal analysis of the subjective questionnaire data, reliability and validity tests were first conducted on the scales measuring three dimensions: spatial presence, immersion, and cultural resonance. The results showed that the overall KMO measure of the scale was 0.842, and Bartlett's test of sphericity reached a significant level ($x^2 = 342.15$, $P < 0.001$), indicating that the data were highly suitable for factor analysis. Principal component analysis extracted three common factors with eigenvalues greater than 1, with a cumulative variance explanation rate of 76.84%. The Cronbach's α coefficients for each dimension ranged from 0.815 to 0.894 (overall $\alpha = 0.887$), all exceeding the critical value of 0.7, indicating that the measurement scales adopted in this study possessed extremely high internal consistency and construct validity.

Using R software, an independent samples t-test was conducted on the subjective ratings of the experimental group (N=30, listening to generative spatial audio) and the control group (N=30, listening to traditional stereo). The specific scores and significance of differences for each perceptual dimension are shown in Table 3.

Table 3. Comparison of differences in subjective perceptual dimensions between the experimental group and the control group (M±SD, N=60)

Dimension	Experimental group (*n*=30)	Control group (*n*=30)	*t*	*P*	Effect size
Spatial presence	4.52\pm 0.48	3.12\pm 0.65	9.482	<0.001	1.03
Immersion	4.38\pm 0.51	3.45\pm 0.58	6.591	<0.001	0.72
Cultural resonance	4.26\pm 0.55	3.58\pm 0.62	4.496	<0.001	0.54

An examination of the experimental results revealed that the group receiving the generative spatialized audio treatment generally outperformed the control group across all indicators. Scores for spatial presence, immersion, and cultural resonance were all markedly higher, while their mean heart rate and skin conductance levels also showed significant increases. These data validate the core concept of this study—that the reconstructed spatial sound field can indeed mobilize greater bodily participation and emotional investment in the audience. According to embodied cognition theory, this implies that spatialized audio more effectively activates the body's sensory information, thereby intensifying the feeling of "being there."

Based on the above research, this paper proposes the concept of "embodied empathy" to describe the mechanism through which spatial audio functions in the process of cross-cultural communication. Embodied empathy refers to a process in which, when situated within an immersive sound field, the audience generates emotional identification with the experiences of others through bodily perception. In facilitating

this process of emotional identification, sound is no longer merely an information carrier but becomes a perceptual medium connecting different cultural subjects. The reconstruction of spatial structure enables the audience to attain a sense of "presence" in the virtual environment, and subsequently approach the performers' experiential state emotionally. Compared with traditional communication models based on symbolic interpretation, embodied empathy emphasizes perceptual priority and experience orientation, and its advantage lies in bypassing the comprehension barriers caused by linguistic and cultural differences.

6. Conclusion

Through in-depth field collection and multimodal data processing, this study has developed a soundscape reconstruction pipeline that takes audio, visual, and spatial information as input, and has verified the feasibility of a paradigmatic pathway in which generative artificial intelligence assists in the reconstruction of ICH soundscapes. In the course of technical practice, this study integrated spatial auditory models such as binaural rendering to carry out spatialized representation and reconstruction of the original polyphonic recordings. After listening to the spatial audio, participants in the experimental group gave significantly higher ratings than the control group with traditional stereo for indicators such as the sense of space and emotional resonance of the folk songs, indicating that the spatial reconstruction conducted in this research indeed enhanced the effectiveness of embodied communication. The spatial reconstruction via generative AI not only restored the lost spatial attributes of ICH sounds, but more importantly, improved cross-cultural communication effectiveness by activating bodily perception. The results confirm that AI plays a key role in the "auditory space" paradigm.

Meanwhile, the study also found that spatial audio, by reconstructing the direction of sound sources, distance relationships, and environmental reflection characteristics, can provide the audience with a spatial perception close to that of the real site. Such perception does not rely on linguistic comprehension, but operates directly through the body's mechanisms of localization and judgment. When audiences experience and understand ICH solely through hearing, their mode of perception shifts from passive reception to active participation, thereby lowering the threshold of cultural understanding. Furthermore, the establishment of this spatial perception not only improves the efficiency of information acquisition but also exerts a very significant influence at the emotional level. Viewers from different cultural backgrounds often find it difficult to establish an emotional connection with unfamiliar musical forms through melodic structure or linguistic content; however, when the sound is presented in a spatialized manner, they can rely on bodily experience to perceive the relationships among the performers, the synergy of the group, and the modes of interaction between people and the environment. This perception-based path of understanding allows emotional resonance to be generated directly through experience, no longer dependent on the accumulation of cultural knowledge. When sound is embedded in three-dimensional space, the audience's understanding of cultural information becomes more multi-dimensional. This finding provides an empirical basis for future research on the digitization and cross-cultural communication of ICH sounds.

In view of the above, future research in related fields should be grounded in the core perspective of spatial computing and centered on the core goal of the generative reconstruction of ICH soundscapes. Efforts should be made to expand the multi-scene and multi-context soundscape datasets for Gyalrong Tibetan polyphonic folk songs, thereby enhancing the generalization capacity and cultural adaptability of the

spatial computing model; to optimize personalized spatial rendering solutions and research deep-learning-integrated generative reconstruction algorithms, so as to refine the pathway for spatialized restoration of ICH soundscapes; and to construct a full-dimensional embodied communication evaluation system covering both short-term immersive experience and long-term cultural identity. Ultimately, through information technology, the contemporary revitalization of ICH soundscapes will be empowered, providing solid theoretical and technical support for the cross-temporal and cross-spatial communication of intangible cultural heritage.

Disclosure statement

The author declares no conflict of interest.

References

[1] Fang L, Lu R, 2023, Chinese Art Anthropology. Minzu Yishu, 2023(6): 152–168.

[2] Gong C, 2024, Reconstruction of Auditory Culture under Traditional Aesthetics. Xiju Shijie (Xiabanyue), 2024(4): 94–96.

[3] Hou J, 2024, Research on the Inheritance Path of Music Intangible Cultural Heritage under the Background of Cultural and Tourism Integration: A Case Study of Qiaoqi Multi-voice Folk Songs. Minzu Yinyue, 2024(3): 37–39.

[4] Hu F, 2025, AI-curated Soundscapes: Spatial Reconfiguration of Intangible Cultural Heritage Music in urban China. GeoJournal, 90(6): 273.

[5] Li X, 2024, Research on the Development Status of Qiaoqi Multi-voice Folk Songs. Yishu Pingjian, 2024(2): 1–6.

[6] Liang W, 2025, Research on the Historical Origin and Cultural Inheritance Path of Dong Grand Song in Guangxi. Wenhua Xuekan, 2025(2): 85–88.

[7] Wang C, Wang Q, 2025, Cultural Penetration and Local Protection of Multi-voice Folk Songs in Qiaoqi Tibetan Township. Xiju Zhijia, 2025(23): 82–84.

[8] Zhang Y, 2025, Research on the Concrete Reconstruction of Sound Scenes by 3D Animation Technology in Music Documentaries. Liuxing Gequ, 2025(9): 40–42.

[9] Zhao S, Tan H, 2025, Soundscape. Minzu Yishu, 2025(5): 151–163.