Digital Intelligence PRESS

Journal of Digital Frontier

https://doi.org/10.67541/jdf2603

Research Article **Volume 1, Issue 1 (July 2026)** *Open Access*

Computational Modeling of Soundscape and Visual Imagery in Beijing Old City Communities Based on Multi-modal Data

Ziyan Shen[1], Xuanyi Qi[2], Mengyu Liu[3]*

[1]*College of mechanical and electrical engineering, Beijing University of Chemical Technology, Chaoyang District, Beijing, China*
[2]*Department of Art and Design, Beijing City University, Shunyi District, Beijing, China*
[3]*Digital Media Art Specialty, Department of Art and Design, Beijing City University, Shunyi District, Beijing, China*

***Corresponding author:** Mengyu Liu. Email: daimou971231@naver.com

Abstract

As an important carrier of the spirit of place, soundscape lacks systematic quantitative evaluation tools. The fundamental bottleneck lies in the lack of an accurate cross-modal alignment calculation framework between soundscape temporal sequence signals and visual image spatial semantics. In this paper, we propose a joint embedding model driven by cross-modal contrastive learning to realize the dynamic coupling of acoustic physical-perceptual features and visual global-local semantic features. We introduce a cross-modal attention forgetting gate to suppress modality-specific noise, design a hierarchical contrastive loss function to embed spatial structure with both instance-level and category-level constraints, and construct a spatio-temporal adaptive weight regulator to realize spatially differentiated modulation of the modality fusion coefficient. Based on 8,432 groups of measured multimodal data (including day-night differences) from four typical community types in Beijing's old city, our experiments show that the proposed model achieves a cross-modal retrieval mAP of 76.82%, a perception score prediction RMSE as low as 0.318, and an intraclass correlation coefficient of 0.824 with manual scores, all of which are significantly better than those of the comparison methods ($p < 0.001$). Ablation experiments verify the independent contributions of the three innovations. Feature response analysis reveals that the partial correlation coefficient between soundscape roughness and visual building density exceeds 0.78, indicating a deep sensory coordination mechanism between auditory texture density and visual spatial envelopment perception. The proposed model provides an effective computational tool for quantitative diagnosis of multi-sensory quality in old city communities.

Keywords

Soundscape; Visual imagery; Multimodal learning; Cross-modal correlation; Beijing old city community

Received: 31 May 2026; Revised: 15 Jul 2026; Accepted: 23 Jul 2026; Published: 06 Aug 2026

1. INTRODUCTION

As an important carrier of historical and cultural heritage, the protection and renewal of Beijing's old city communities have long focused on the control and restoration of architectural style, street texture, and visual landscape [1],[2]. However, as an important part of the spirit of place, soundscape has long been absent from the protection practices of old city communities. Different from the logic of sound environment management dominated by noise control in modern cities, the auditory experience of old city communities carries richer cultural memory and daily life information—pigeon whistles in

the morning, hawking calls in the afternoon, and neighborhood conversations in the evening [3]. These sounds, with their distinct regional characteristics, constitute a unique urban auditory landscape. However, with the iteration of commercial formats and the updating of infrastructure in the old city, the traditional soundscape is being rapidly eroded by new sound sources such as loudspeakers, the hum of outdoor air-conditioning units, and electric vehicle alarms [4],[5],[6]. Nevertheless, the current protection policy lacks systematic indicators and control measures for auditory quality, resulting in the one-sided situation of "only focusing on its form and not listening to its sound." It has therefore become a research proposition of both theoretical value and practical urgency to establish a computational model that associates soundscape with visual imagery and provides a quantitative tool for multi-sensory quality assessment in old city communities.

From a computational perspective, existing research on soundscape and visual imagery has respectively formed relatively complete methodological systems—soundscape research relies on psychoacoustic indicators and signal processing techniques to describe auditory perception, while visual imagery research uses deep convolutional networks and scene understanding models to analyze spatial semantics—but the two mostly develop in parallel as independent tracks [7],[8]. Cross-modal association analysis is still in its infancy. The bottlenecks of existing cross-modal work are reflected in two aspects. First, the mismatch of spatio-temporal granularity fundamentally limits the effective alignment of soundscape and visual information [9],[10]. Soundscape is a continuously evolving temporal signal with millisecond-level temporal sensitivity, whereas visual imagery is usually extracted from still images and reflects a snapshot of the spatial scene at the sampling time. Traditional methods often simply align the two on a coarse time scale, ignoring the asynchronous relationship between acoustic events and visual scenes within a few seconds, and thus fail to establish accurate semantic mappings [11],[12]. Second, modality alignment lacks a computational framework with physical constraints and perceptual basis. Existing methods mostly adopt naive feature stitching or shallow attention mechanisms, which fail to fully consider that not all components between soundscape and visual imagery are associated—for example, lighting changes in vision and wind noise in soundscape are modality-specific interference [13],[14]. If indiscriminate global fusion is performed, it will weaken the shared semantics that truly reflect the commonality of places. As a result, the model accuracy in cross-modal retrieval and perception prediction tasks remains far from practical levels.

To solve the above bottlenecks, this paper proposes a computational framework for joint embedding of soundscape and visual image driven by cross-modal contrast learning, and its core contributions lie in three aspects [15]. First, at the level of feature extraction, the physical-perceptual dual-domain soundscape feature pool and hierarchical visual semantic feature extraction pipeline are designed to completely describe the soundscape from the two dimensions of signal attributes and subjective auditory experience, and at the same time, the visual image is analyzed from two granularities of global scene and local element, providing high-quality heterogeneous input for subsequent cross-modal modeling. Secondly, at the model architecture level, the cross-modal attention forgetting gate is introduced to dynamically filter out the modality-specific noise, the hierarchical contrast loss function is designed to simultaneously constrain the geometric structure of the embedding space at the instance and category levels, and the spatio-temporal adaptive modality weight regulator is proposed to realize the spatial differentiation regulation of the fusion coefficient. The three innovations work together to improve the discriminability and robustness of the joint representation. Thirdly, at the level of experimental verification, this paper constructs a large-scale measured data set covering four typical communities in the old city of Beijing, completes the closed-loop evaluation from the three dimensions of retrieval accuracy, regression error and human-machine consistency, and quantitatively reveals the

strong coupling relationship between soundscape roughness and visual building density, which provides an interpretable computational basis for the quantitative diagnosis of multi-sensory quality of the old city community.

The technical route of this paper is as follows: first, sound pressure level, panoramic images, and GPS/IMU positioning data are synchronously collected using a mobile sensor array; after feature extraction and dynamic time warping alignment, standardized soundscape and visual feature sequences are constructed. Second, with the dual-stream encoder as the benchmark framework, we successively embed the cross-modal attention forgetting gate, hierarchical contrastive loss, and spatio-temporal adaptive weight regulator to train the cross-modal joint embedding model. Finally, model performance is comprehensively evaluated through comparative experiments, ablation experiments, and subjective perception verification, and the coupling relationships of key features are revealed based on significance tests and visual analysis. Section 1 expounds the research background and problem boundaries; Section 2 details the data acquisition scheme and joint feature extraction method, Section 3 systematically presents the core model architecture and algorithm innovation, Section 4 verifies the effectiveness of the model and the contribution of each module through the complete experimental chain, and Section 5 summarizes the conclusions and points out limitations and future directions. A progressive logic from data to features and from model to verification is formed among the sections. Within each section, a self-consistent closed-loop of methodology and result analysis is proposed to support the overall argument.

2. MULTI-MODAL DATA ACQUISITION AND JOINT FEATURE EXTRACTION METHOD

This section focuses on the engineering acquisition strategy for raw soundscape and visual image data from Beijing's old city communities, and on the high-quality representation method for heterogeneous features. The quality of data collection and feature extraction directly determines the upper limit of subsequent cross-modal correlation modeling [16]. Therefore, targeted design is carried out in this section in sensor layout, sampling protocol, feature engineering and spatio-temporal alignment, so as to adapt to the special environmental constraints of high density, narrow streets and strong spatial heterogeneity of old city communities.

The mobile sensor array acquisition system is composed of three parts: portable multi-channel acoustic recorder, panoramic visual acquisition unit and high-precision integrated navigation and positioning module. G.R.A.S. 46AE 1/2 inch free-field microphone was used for acoustic recording, with NI 9234 data acquisition card. The sampling frequency was set to 44.1 kHz, the quantization bit depth was 24 bit, and the frequency response range was from 20 Hz to 20 kHz, which met the requirements of full frequency band recording of human hearing. The panoramic vision unit is installed with four Flir Blackfly S industrial cameras in an orthogonal way, with a fisheye lens (field of view Angle of 190°). The hardware synchronization trigger realizes simultaneous exposure of four channels of images, and the panoramic image in isometric column projection format is output after splicing, with a resolution of 4096×2048 pixels. The positioning and attitude module adopts NovAtel SPAN-IGM-S1 integrated navigation system, which integrates GPS L1/L2 dual-frequency signals with IMU three-axis gyroscope and accelerometer data. The post-processing differential positioning accuracy can reach centimeter level, and the heading Angle accuracy is better than 0.1°. The three subsystems are synchronized through the same embedded industrial computer (supporting IEEE 1588 precise time protocol) to ensure that each frame of visual image, each second acoustic segment and the

corresponding spatial coordinates have a unified timestamp reference [17],[18]. The acquisition strategy is carried out along the typical streets and lanes of the old city community, and the moving speed is controlled between 0.8m /s and 1.2m /s. It takes into account the stability of sound field and motion blur suppression of panoramic image. Data packaging and storage are automatically triggered every 5 meters, and auxiliary meteorological parameters such as ambient temperature and wind speed are recorded for subsequent data screening.

Soundscape feature extraction is not simply a pile-up of acoustic physical quantities, but rather the construction of a multi-dimensional feature pool covering the low-level attributes of signals and the high-level attributes of auditory perception, so as to fully describe the complex mixing characteristics of sound source composition of old city communities (such as human voice, traffic sound, bird song, construction noise, etc.) [19]. The collected time-domain acoustic signal is set as $x(t)$, and the i-th frame signal $x_i(n)$ is obtained after Hamming window framing and windowing. The frame length is 1024 sampling points, and the frame shift is 512 sampling points. Firstly, the basic physical features are extracted from the time-frequency domain. The calculation process of Mel frequency cepstral coefficient (MFCC) is as follows: The short-time Fourier transform is applied to $x_i(n)$ to obtain the power spectrum $|X_i(k)|^2$, and the linear frequency is mapped to the Mel frequency scale through a set of triangular Mel filter banks $H_m(k)$ (the number of filters $M = 26$). After taking the logarithmic energy, the 12-dimensional static MFCC coefficients are obtained by discrete cosine transform. Together with the first-order and second-order differences, a 36-dimensional dynamic feature vector is formed, and its calculation formula is as follows:

$$c_j = \sum_{m=1}^{M} \log\left(\sum_{k=1}^{K} |X_i(k)|^2 H_m(k)\right) \cos\left(j\left(m-\frac{1}{2}\right)\frac{\pi}{M}\right), j = 1,2,\ldots,12 \tag{1}$$

Where K is the number of FFT points and j is the order of cepstral coefficients. spectral centroid measures the center of gravity of spectral energy and reflects the brightness attribute, which is defined as:

$$SC_i = \frac{\sum_{k=1}^{K} f_k |X_i(k)|^2}{\sum_{k=1}^{K} |X_i(k)|^2} \tag{2}$$

Where f_k is the frequency value corresponding to the k frequency bin. loudness fluctuation represents the time-varying intensity of sound pressure level. After calculating the instantaneous loudness $N'(z,t)$ in a specific frequency band according to the Zwicker model, the coefficient of variation of loudness with time is calculated [20].

On top of physical features, perceptual psychoacoustic indicators are further introduced to bridge the gap between physical signals and subjective auditory experience. Sharpness reflects the proportion of high-frequency components in the total energy, and its calculation is based on the weighted integral of a specific loudness spectrum $N'(z)$. The typical weight function $g(z)$ increases exponentially with the critical frequency band rate z:

$$S = 0.11 \cdot \frac{\int_0^{24\text{Bark}} N'(z) g(z)\, dz}{\int_0^{24\text{Bark}} N'(z)\, dz} \tag{3}$$

In the formula, the integral interval covers 24 Bark critical frequency bands corresponding to the whole audible frequency band of the human ear, and the unit of sharpness is acum. Roughness is used

to quantify the auditory discomfort caused by fast amplitude modulation (modulation frequency between 20 Hz and 300 Hz) in sound signals. The estimation is based on the nonlinear function relationship between the modulation depthΔL_z and the modulation frequency f_{mod} for each frequency band. Tonality reflects the prominence of pure tone component relative to noise component in the signal, which is obtained by calculating the significance of deviation between spectral peak of each frequency band and the average value of adjacent spectrum. The above seven types of features (36-dimensional MFCC, 1-dimensional spectral centroid, 1-dimensional loudness fluctuation coefficient of variation, 1-dimensional sharpness, 1-dimensional roughness, and 1-dimensional sound schedule) together constitute the 41-dimensional original feature vector of each frame acoustic signal, and then take the statistics (mean, standard deviation, skewness, kurtosis) of each frame feature in a 5-second window along the time series. Finally, the soundscape corresponding to each sampling point was compressed into a fixed length representation vector of 164 dimensions (41 features ×4 statistics).

Feature extraction of visual image follows a hierarchical and progressive strategy to capture the potential modulation effect of visual environment on auditory perception from two granularities of global scene semantics and local visual elements respectively. For the panoramic image collected simultaneously at each sampling point, it is first cut into six perspective projection planes to eliminate the adverse effect of spherical distortion on the convolution operation. The global scene semantic branch uses ResNet-50 pre-trained on ImageNet-21K as the backbone network to extract 2048-dimensional spatial feature maps, which are then flattened and input to a Transformer encoder module to capture scene context dependencies in the global receptive field [21]. The Transformer encoder is composed of four layers of multi-head self-attention (the number of attention heads is set to 8) and feedforward network. Its core self-attention mechanism is described as follows:

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

Where Q, K and V are obtained from the input features by linear projection respectively, and d_k is the dimension of the key vector, where $d_k = 256$. The global features output by Transformer encoder are reduced to 1024 dimensions after average pooling, and the vector contains higher-order scene semantics such as “narrowness of hutong,” “openness of square,” and “building density”. The local visual element branch explicitly quantifies the fine-grained environmental elements unique to the old city community. The semantic masks of roof shape, gable style, and door-window proportion are obtained by semantic segmentation of the image (using DeepLabV3+ network, fine-tuning on the self-annotated dataset of Beijing old City buildings), and then compressed into 256-dimensional style vector by convolutional autoencoder. The greening rate is calculated by the proportion of vegetation pixels in the effective area of the panorama in the segmentation results, which is denoted as G_r. The skyline openness is defined as the ratio of the Angle range of the sky region in the panoramic image to the total Angle of 360°, which is measured after the sky region is extracted by the joint method of color threshold and edge detection [22]. The final output dimension of the local visual branch is 512 (256 dimensions of style vector + 128 dimensions of location coding + 128 dimensions of green rate and openness concatenation).

In the process of mobile acquisition, the spatial interval between sampling points is constant but the traveling speed fluctuates slightly, resulting in nonlinear offset between acoustic feature time series and visual feature series on the time axis, and direct splicing according to the acquisition sequence will introduce feature misalignment error. Therefore, dynamic time warping (DTW) algorithm is introduced to preprocess the spatio-temporal alignment of the two types of feature sequences. Let the soundscape

feature sequence be $A = \{a_1, a_2, \ldots, a_M\}$, visual characteristic sequence for $B = \{b_1, b_2, \ldots, b_N\}$, where each a_i and b_j are the fixed-length feature vectors extracted above. Sequence lengths M and N correspond to the number of valid data frames in the same acquisition period. DTW finds the optimal alignment path by constructing an $M \times N$ cumulative distance matrix D, and its recurrence relation is as follows:

$$D(i,j) = d(a_i, b_j) + \min\{D(i-1,j), D(i,j-1), D(i-1,j-1)\} \quad (5)$$

The initial conditions are set as $D(0,0) = 0$, $D(i,0) = \infty$ $(i > 0)$, and $D(0,j) = \infty$ $(j > 0)$. Where $d(a_i, b_j)$ is the Euclidean distance between two feature vectors, that is:

$$d(a_i, b_j) = \sqrt{\sum_{p=1}^{P} (a_{i,p} - b_{j,p})^2} \quad (6)$$

Here, P is the total dimension of feature vectors. In order to ensure the balanced contribution of acoustic and visual features in distance calculation, independent Z-score standardization is carried out on each feature dimension in advance. Through dynamic programming to solve the cumulative distance minimum path $\pi = \{\pi_1, \pi_2, \ldots, \pi_L\}$, Including $\pi_l = (i_l, j_l)$ said it would first i_l a sound landscape characteristics and the first j_l a visual feature matching alignment. The path must satisfy boundary conditions ($\pi_1 = (1,1)$, $\pi_L = (M,N)$), monotonicity ($i_{l+1} \geq i_l$, $j_{l+1} \geq j_l$) and continuity (adjacent path point step length of at most 1). After the alignment is completed, the paired audio-visual features are resampled to a unified time reference to form a one-to-one corresponding soundscene representation vector and visual image representation vector at each spatial sampling point, which are used as standard input data for subsequent cross-modal computational modeling. It is worth pointing out that, on the basis of standard DTW, this paper adds a window constraint (Sakoe-Chiba band, the bandwidth is 10% of the sequence length) to suppress excessive distortion and ensure that the alignment path is only active within a physically reasonable timing deviation range, so as to eliminate the spurious misalignment introduced by acquisition jitter while retaining the true delay relationship.

3. COMPUTATIONAL MODEL OF CROSS-MODAL CORRELATION BETWEEN SOUNDSCAPE AND VISUAL IMAGERY

In this section, a nonlinear mapping framework from heterogeneous multi-modal features to joint semantic space is constructed. The core challenge is that there is a natural representation gap between the one-dimensional temporal attributes of soundscape and the two-dimensional spatial attributes of visual image, and the two modes in the old city community environment contain a large number of interference information unrelated to each other (such as wind noise in acoustic mode and lighting change in visual mode). Direct fusion will seriously weaken the activation strength of shared semantics. Therefore, based on the benchmark structure of dual-stream encoder, this paper introduces three progressive innovations, namely, cross-modal attention forgetting gate, hierarchical contrast loss function and spatio-temporal adaptive weight regulator, to form a complete set of cross-modal correlation calculation model.

The benchmark model adopts dual-stream encoder architecture to perform independent high-order abstraction of soundscape features and visual image features [23]. The soundscape encoder takes the

164-dimensional soundscape representation vector $s \in \mathbb{R}^{164}$ extracted in Section 2 as the input, and the timing coding backbone is formed by stacking three layers of gated recurrent units (GRU). The number of hidden units in each layer is 256, 512 and 256 in turn. Residual connections are introduced between layers to alleviate gradient degradation. The core update mechanism of GRU is as follows:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{7}$$

$$R_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{8}$$

$$\tilde{H}_t = \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h) \tag{9}$$

$$H_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{10}$$

Let x_t be the input feature vector at time t, h_{t-1} for the first time the hidden state, z_t and r_t update and reset the door vector, respectively, $\sigma(\cdot)$ is the Sigmoid activation function, $\odot$ represents the element-by-element product, and W, U and b are learnable parameter matrices and bias terms. The final hidden state of GRU output is mapped to the soundscape embedding vector $e_s \in \mathbb{R}^d$ with dimension $d = 512$ by the linear projection layer. The visual encoder takes the visual image feature vector $v \in \mathbb{R}^{1536}$ (composed of 1024-dimensional global scene semantics and 512-dimensional local visual elements) output in Section 2 as input, and adopts the four-layer Transformer encoder structure. Each layer contains multi-head self-attention (8 heads) and feedforward network (hidden dimension 2048), and LayerNorm is used to connect the residual between layers. Multi-head self-attention performs parallel projection of input features:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_8)W^O \tag{11}$$

$$\text{Head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{12}$$

Where Q, K, V are all obtained from the same input sequence through different linear transformations. $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{1536\times192}$. $W^O \in \mathbb{R}^{1536\times1536}$. After four layers of coding, the output corresponding to the [CLS] token is taken as the overall visual representation, and then the visual embedding vector $e_v \in \mathbb{R}^d$ is obtained through the projection layer, which maintains the same dimensional space as the soundscene embedding. In the benchmark model, e_s and e_v are concatenated to output the joint association score through the two-layer fully connected network.

Based on the benchmark dual-stream structure, this paper firstly proposes the Cross-modal Attention Forget Gate (CM-AFG), which aims to solve the noise amplification problem caused by "excessive attention" in the traditional cross-modal attention mechanism. The soundscape of the old city community contains a large number of transient sound events that have nothing to do with the visual scene (such as the distant car whistle and the friction sound of the kite line), and there are also dynamic light and shadow changes in the visual mode that have nothing to do with the hearing. If these interference components participate in the cross-modal interaction without inhibition, the real co-occurrence semantic signals will be diluted. The core idea of CM-AFG is to introduce a learnable forgetting gate vector to “gate” the source modality features the source modal features before calculating the cross-modal attention weight. Specifically, when the soundscape feature e_s queries attention from the visual feature e_v, the forgetting gating coefficient is first calculated:

$$g_{s\to v} = \sigma\left(W_g^s e_s + U_g^s e_v + b_g^s\right) \tag{13}$$

Where $W_g^s, U_g^s \in \mathbb{R}^{d\times d}$, $b_g^s \in \mathbb{R}^d$ are learnable parameters, Each element in $g_{s\to v} \in (0,1)^d$

controls whether the information on the corresponding dimension is allowed to participate in the subsequent attention calculation. Then the soundscene embedding is gated and modulated to obtain the filtered query vector $\tilde{\mathrm{e}}_s = \mathrm{g}_{s\to v} \odot \mathrm{e}_s$. The weight of cross-modal attention is:

$$\alpha_{s\to v} = \mathrm{softmax}\left(\frac{(\tilde{\mathrm{e}}_s \mathrm{W}^Q)(\mathrm{e}_v \mathrm{W}^K)^\top}{\sqrt{d_k}}\right) \tag{14}$$

Type $\mathrm{W}^Q, \mathrm{W}^K \in \mathbb{R}^{d\times d_k}$ for projection matrix (take $d_k = 64$), $\alpha_{s\to v}$ is the d-dimensional attention weight vector. A similar forgetting gating process is applied in reverse to the attention direction from vision to soundscape, and $\mathrm{g}_{v\to s}$ and $\tilde{\mathrm{e}}_v$ are obtained. The design of the gating coefficient enables the model to dynamically judge which feature dimensions have cross-modal correlation according to the current input sample, and the interference dimension whose inhibition gate coefficient tends to 0 is cut out from the attention calculation, thus forcing the model to focus on the part of real semantic co-occurrence in the two modes. CM-AFG does not change the embedding dimension, but only inserts a soft screening mechanism before attention calculation, and obtains significant anti-noise gain with a minimal number of parameters (an increase of about $4d^2$).

After obtaining the embedded representation screened by the forgetting gate, the core link of model training lies in the design of loss function. In this paper, Hierarchical Contrastive Loss (HCL) function is proposed. Based on the instance level alignment of traditional contrast learning, category-level prototype constraints are introduced [24],[25]. In order to cope with the dilemma that there is a large variance in the audio-visual correlation mode of the same scene in the old city community (like the historical reserve), while there may be accidental similarity between different scenes [26]. Let a training batch contain N samples for each sample by sound scene embedded $\mathrm{e}_s^{(i)}$ and visual embedded $\mathrm{e}_v^{(i)}$, There are N positive sample pairs and $N(N-1)$ negative sample pairs in the batch. The instance-level contrastive loss takes the InfoNCE form which encourages positive sample pairs to approach each other and negative sample pairs to move away from each other in the joint embedding space:

$$\mathcal{L}_{\text{inst}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\dfrac{\mathrm{sim}\left(\mathrm{e}_s^{(i)}, \mathrm{e}_v^{(i)}\right)}{\tau}\right)}{\sum_{j=1}^{N} \exp\left(\dfrac{\mathrm{sim}\left(\mathrm{e}_s^{(i)}, \mathrm{e}_v^{(j)}\right)}{\tau}\right)} \tag{15}$$

The $\mathrm{sim}(\mathrm{u},\mathrm{v}) = \mathrm{u}^\top \mathrm{v}/(\| \mathrm{u} \| \| \mathrm{v} \|)$ is the cosine similarity function, and τ is the temperature coefficient (0.07 in this paper). According to the spatial location and functional attributes of the collection points, all samples are pre-divided into C scene categories (such as traditional residential type, mixed commercial and residential type, historical protection type, etc., $C = 5$), and the centroid embedded by all samples in each category is calculated online:

$$\mathrm{c}_s^{(k)} = \frac{1}{|P_k|}\sum_{i\in P_k} \mathrm{e}_s^{(i)}, \mathrm{c}_v^{(k)} = \frac{1}{|P_k|}\sum_{i\in P_k} \mathrm{e}_v^{(i)} \tag{16}$$

Here P_k represents the sample index set belonging to the k class, $|P_k|$ is the number of samples of this category. Category-level contrast loss aligns the centroid of the same soundscene category with that of the corresponding visual category, and at the same time increase the distance between the centroids of different categories:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{C}\sum_{k=1}^{C}\log\frac{\exp\left(\frac{\text{sim}\left(\text{c}_s^{(k)},\text{c}_v^{(k)}\right)}{\tau}\right)}{\sum_{l=1}^{C}\exp\left(\frac{\text{sim}\left(\text{c}_s^{(k)},\text{c}_v^{(l)}\right)}{\tau}\right)} \tag{17}$$

Eventually layered contrast loss of two weighted array $\mathcal{L}_{\text{HCL}} = \lambda\mathcal{L}_{\text{inst}} + (1-\lambda)\mathcal{L}_{\text{cls}}$, The weight coefficient λ is set to 0.7 to retain the dominance of instance level fine-grained alignment, while introducing category level macro constraints. This design enables the model not only to distinguish the subtle acoustic-visual correlation differences between individual samples within the same category, but also to ensure that samples from different scene categories form a separable structure in the embedding space, effectively avoiding the problem of "pushing the similar samples away by mistake" in the instance level comparison learning.

The third innovation addresses the dynamic change of modal contribution caused by spatial heterogeneity within old city communities. In the areas with different community densities, the contribution weight of soundscape and vision to the overall image should be different -- in the open square area, vision is more dominant; In narrow hutong, the spatial indication of soundscape is more prominent. Therefore, a Spatio-Temporal Adaptive Modality Weighting (STAMW) is designed to dynamically calculate the fusion coefficient according to the spatial-temporal density of GPS coordinates of sampling points [27]. First for each sampling point location $\text{p}_i = (\text{lat}_i, \text{lon}_i)$, calculate the local neighborhood density ρ_i :

$$\rho_i = \frac{1}{\pi r^2}\sum_{j=1}^{N}\mathbb{1}\left(\|\ \text{p}_i - \text{p}_j\ \|_2 \leq r\right) \tag{18}$$

Where r is the spatial bandwidth (25 meters), $\mathbb{1}(\cdot)$ is the indicator function, and the unit of ρ_i is sample point/square meter, reflecting the sampling density of this area, that is, the spatial complexity. At the same time, the time stamp t_i is introduced to calculate the time factor$\eta_i = \sin(\pi \cdot t_i/T)$, T is the interval between day and night (12 hours), and $\eta_i \in [0,1]$ represents the time difference between day and night. Then the density ρ_i and the time period factor η_i are input into a lightweight multi-layer perceptron (hidden layer 32 dimensions) to output the modal weight:

$$w_i^s = \frac{\exp(f_{\text{MLP}}([\rho_i,\eta_i])_s)}{\exp(f_{\text{MLP}}([\rho_i,\eta_i])_s) + \exp(f_{\text{MLP}}([\rho_i,\eta_i])_v)} \tag{19}$$

$$w_i^v = 1 - w_i^s \tag{20}$$

Where w_i^s and w_i^v are the fusion weights of soundscene and vision at the sampling point respectively, and the sum of the two is always 1. The final joint embedding vector is obtained by weighted splicing:

$$\text{e}_{\text{joint}} = \text{Concat}\left(w_i^s \cdot \text{e}_s^{(i)},\ w_i^v \cdot \text{e}_v^{(i)}\right) \tag{21}$$

e_{joint} is the feature representation ultimately used for downstream association score prediction. The regulator design introduces no additional hyperparameters, and all parameters are trained end-to-end with the main model, so that the model has the adaptability to automatically adjust the modal contribution degree according to the actual spatio-temporal properties of the collection environment.

We used the AdamW optimizer for model training, the initial learning rate was set to 5×10^{-4}, and the cosine annealing attenuation strategy was dynamically adjusted. Aiming at the joint training of dual-stream encoder and three innovation modules, the we introduce a gradient clipping strategy to suppress gradient explosion:

$$\text{If } \| \mathrm{g} \|_2 > \theta, \mathrm{g} \leftarrow \theta \cdot \frac{\mathrm{g}}{\| \mathrm{g} \|_2} \tag{22}$$

Where g is the gradient splicing vector of all trainable parameters, and the clipping threshold θ is 1.0. The early stop mechanism monitors the mean accuracy (mAP) of cross-modal retrieval on the validation set. If there is no improvement in 15 consecutive training rounds, the training is terminated and the optimal model parameters are rolled back. For the convergence discussion, Model of $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{HCL}} + \gamma \| \mathrm{W} \|_2^2$ ($\gamma = 10^{-4}$ is L2 regularization coefficient) decreases rapidly in the initial stage of training and enters a stable period after about 80 rounds. Because we use exponential moving average (EMA) to smooth category fluctuations when the category centroid in hierarchical contrast loss is updated in each round, the training curve shows no signs of shock and divergence. Among the three innovation modules, the forgetting gate parameter introduced by CM-AFG is initialized to a bias close to 1 (b_g is initially positive), so that the model is similar to standard attention in the early stage of training, and the gating value distribution is gradually acquired with the progress of training. This initialization strategy effectively avoids the cold start problem. The total number of model parameters is about 28 million. A single training epoch on an NVIDIA RTX 4090 takes about 12 minutes, which meets the computational efficiency requirements for practical applications.

4. EXPERIMENTAL DESIGN AND RESULT VERIFICATION

In order to comprehensively evaluate the effectiveness of the cross-modal association calculation model proposed in this paper, a complete closed-loop experimental chain from data construction, quantitative comparison, and module ablation to subjective perception verification is designed in this section. The experiment covers four typical community types in the old city of Beijing. Based on large-scale measured data, the model performance is strictly tested from multiple dimensions such as retrieval accuracy, regression error and man-machine consistency. All the experimental results are verified by statistical significance to ensure the reliability of the conclusion.

The experimental data were collected in the core area of the old city of Beijing, and four types of typical communities with distinct spatial morphology and functional differences are selected: Historical protection area (represented by the surrounding area of Shichahai, with complete hutong texture and concentration of traditional courtyards), mixed commercial and residential area (represented by the main street of Nanluoguxiang, with dense commercial forms and noisy people flow), hutong micro-renovation area (represented by the Baitasi area, which has been gradually renovated and improved in space quality in recent years), traditional residential area (represented by the inside of Anding Men, The daily living function is dominant and the environment is relatively quiet). Collection routes were set up along the main streets and alleys of each type of community, with the length of each route being about 1.2 to 1.8 kilometers. The mobile collection speed was uniformly controlled at 1.0 m/s, and an effective sampling point was recorded every 5 meters. The collection time covers two periods of day (9:00-11:30, 14:00-17:00) and night (19:00-21:30) to capture the systematic differences of day and night soundscape and visual imagery. After feature extraction and DTW spatio-temporal alignment using the method in Section 2, a total of 8,432 valid sample pairs are obtained, including 1987 groups of historical

conservation areas, 2156 groups of mixed commercial and residential areas, 2143 groups of hutong micro-renewal areas, and 2146 groups of traditional residential areas. According to the ratio of 7:1.5:1.5, the samples were randomly divided into training set (5902 groups), validation set (1265 groups) and test set (1265 groups) to ensure that the proportion of all kinds of samples was balanced in the division. All features were independently Z-score standardized before input to the model. Standardization parameters were computed only from the training set and then applied to the validation and test sets.

We used three core evaluation metrics to quantify model performance from different dimensions. Cosine similarity measures the alignment degree of soundscape embedding and visual embedding output by the model in the joint space. For the i-th test sample pair, it is defined as:

$$\mathrm{CosSim}_i = \frac{\mathrm{e}_s^{(i)} \cdot \mathrm{e}_v^{(i)}}{\| \mathrm{e}_s^{(i)} \|_2 \| \mathrm{e}_v^{(i)} \|_2} \tag{23}$$

The reported value is the average cosine similarity over the entire test set $\overline{\mathrm{CosSim}} = \frac{1}{N_{\mathrm{test}}} \sum_{i=1}^{N_{\mathrm{test}}} \mathrm{CosSim}_i$, The closer the value is to 1, the better the semantic alignment between modes is. The mean accuracy of cross-modal retrieval (mAP) is used to evaluate the performance of the model in bidirectional retrieval tasks, that is, the ranking accuracy of the true matching samples in the returned results when querying visual images with soundscape (or vice versa). The average accuracy (AP) is calculated for each query sample, and then the average is taken for all queries:

$$\mathrm{mAP} = \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{R_q} \sum_{k=1}^{K} \mathrm{Precision}_q(k) \cdot \mathrm{Rel}_q(k) \tag{24}$$

Where Q is the total number of queries, R_q is the total number of relevant samples for the q-th query, K is the length of the return list (50 in this paper), $\mathrm{Precision}_q(k)$ is the precision in the first k returned results, $\mathrm{Rel}_q(k)$ is a binary label indicating whether the k-th result is relevant (1 is relevant, 0 is irrelevant). RMSE of perception score prediction is used to quantify how well the model approximates subjective joint audio-visual perception. The soundscape and panoramic images collected at each sampling point were submitted to five pre-trained raters, and the joint image score $\hat{y}_i$ was given according to the five-point Likert scale from “very congruent” to “very coordinated”. The average of the five people was taken as the real perception score of this point. The model outputs the prediction score y_i for $\mathrm{e}_{\mathrm{joint}}^{(i)}$ through the linear regression head, and the RMSE is calculated as follows [28],[29],[30],[31]:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_{\mathrm{test}}} \sum_{i=1}^{N_{\mathrm{test}}} (y_i - \hat{y}_i)^2} \tag{25}$$

The smaller the index is, the closer the joint representation of the model is to people's real subjective experience.

Table 1 summarizes the performance of the three indicators of each comparison method on the test set. The methods used to participate in the comparison include: Single-mode acoustic reference (using only the soundscape encoder, with the visual part zeroed), single-mode visual reference (using only the visual encoder, with the soundscape part zeroed), simple Concatenation fusion (Concatenation, with the

dual-stream output directly cascading followed by the full connection layer), multi-mode bilinear pooling (MFB, Outer product compression fusion of dual-stream features), standard cross-modal attention (without forgetting gate mechanism), and the complete model of this paper (Ours). It can be clearly observed from Table 1 that the mAP of the single-mode acoustic reference is only 39.27%, indicating that it is difficult to uniquely identify the visual scene only by acoustic information. Although mAP of single-mode visual benchmark increased to 51.83%, RMSE of perception prediction was as high as 0.521, reflecting that there is still a nonlinear gap between visual features and subjective joint feelings. In the classical multi-modal fusion method, MFB is better than simple Mosaic due to its second-order interaction ability (mAP 62.35% vs. 58.91%), while the standard Cross-attention further pushes mAP to 68.14% by virtue of two-way attention mechanism, which verifies the necessity of cross-modal interaction. The complete model in this paper achieves the optimal results in the three indicators: mAP reaches 76.82%, which is 8.68 percentage points higher than the Cross-attention of the optimal comparison method; The average cosine similarity is 0.847, which is 0.056 higher than 0.791 of Cross-attention. RMSE decreased to 0.318, 17.4% lower than 0.385 of Cross-attention. This comprehensive advantage shows that CM-AFG effectively eliminates the interference of irrelevant noise between modes, HCL simultaneously constrains the embedding space at both instance and category levels to make the retrieval ranking more accurate, and STAMW dynamically allot the modal weight according to spatial density to avoid the under-adaptation problem of static fusion in heterogeneous environments.

Table 1. Performance comparison of different methods on the test set

The Method	Average cosine similarity	mAP (%)	Perceived RMSE
Single-modal acoustic baseline	0.582	39.27	0.647
Single-modal visual baseline	0.674	51.83	0.521
Concatenation	0.703	58.91	0.472
MFB	0.735	62.35	0.443
Cross-attention	0.791	68.14	0.385
Ours (full model)	0.847	76.82	0.318

In order to verify the actual contributions of the three innovations in this paper, four groups of ablation experiments are designed: Cross-modal attention forgetting gate (-CM-AFG) was removed, hierarchical contrast loss was removed but only instance level InfoNCE was retained (-HCL), spatio-temporal adaptive mode weight controller was removed and fixed equal weight fusion was adopted (-STAMW), and all three innovations were removed simultaneously and only benchmark dual-stream + instance contrast loss was retained (-All). Each variant is run under the same training configuration, and the results are shown in Table 2. After removing CM-AFG, mAP decreases by 3.91 percentage points and RMSE increases by 0.034, which proves that the forgetting gate plays an irreplaceable role in noise suppression, especially in the environment with frequent acoustic interference in the mixed commercial and residential areas. After removing HCL, mAP decreases by 4.63 percentage points, which is the largest among the three, indicating that the category-level prototype constraint makes the most critical contribution to improving the quality of retrieval ranking. After removing STAMW, the mAP decreases by 2.74 percentage points, and the decrease is the largest in the historical protected areas and the mixed

commercial and residential areas (by 1.8% and 4.1% respectively), which indirectly proves that STAMW plays a more prominent role in the weight adjustment in the areas with strong spatial heterogeneity. When all the innovations are removed at the same time, mAP drops to 61.37%, which is still lower than the MFB method (62.35%), indicating that the three innovations are not simply superimposed, but work together.

Table 2. Comparison of the performance of each variant in the ablation experiment

Variant configuration	Average cosine similarity	mAP (%)	Perceived RMSE
-All (benchmark + instance-level contrast only)	0.694	61.37	0.456
-CM-AFG	0.809	72.91	0.352
-HCL	0.801	72.19	0.365
-STAMW	0.821	74.08	0.337
Ours (full model)	0.847	76.82	0.318

Figure 1 presents the fine-grained mAP performance on the four community types. The bar chart shows the retrieval accuracy differences of each method on different types of communities. The complete model keeps the lead in the four types of communities, but the improvement in different communities shows gradient differences: in the mixed commercial and residential areas, mAP reaches 79.34% (10.27 percentage points higher than Cross-attention), and the improvement is the largest, which confirms that the anti-noise effect of the forgotten gate is the most significant in the area with the most serious acoustic interference. In the traditional residential area, mAP was 74.51% (6.55 percentage points higher than Cross-attention), which showed the smallest increase but still had practical significance. The historical protection area and hutong micro-renewal area reached 77.28% and 76.15% respectively.

It is worth noting that the single-modal visual baseline achieves 54.67% mAP in the historical reserve, while the single-modal acoustic reference is only 36.82% in the MAP of the historical reserve, indicating that the historical reserve is dominated by the high identifiability of the visual form, and the acoustic information is relatively homogeneous. However, in the commercial-residential mixed area, the gap between the gap between the acoustic baseline and the visual baseline narrows to 8.16 percentage points (42.53% vs. 50.69%), reflecting that the gain of acoustic information in scene recognition in this area is more obvious, which also explains the reason why STAMW has a larger adjustment range in this area.

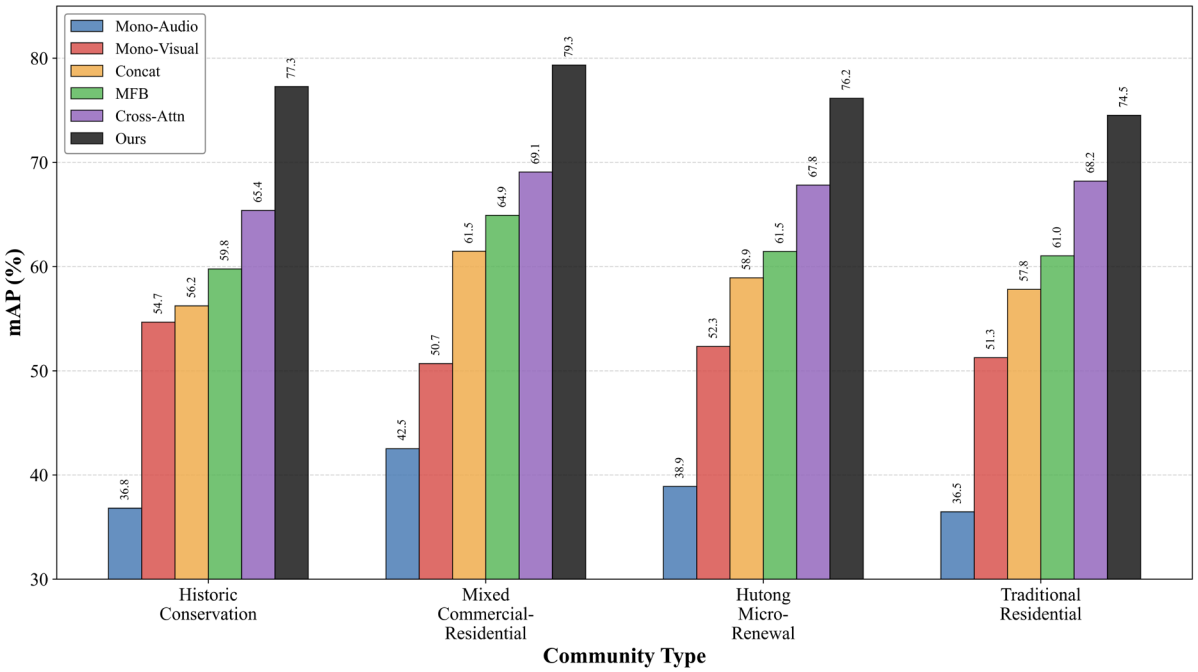

Figure 1. Bar graph of cross-modal retrieval mAP comparison of each method on the four types of communities

In order to further investigate the performance differences of the model in different perceptual intensity intervals, Figure 2 shows scatter density plots of predicted versus ground-truth perception scores. The predictions of the full model cluster tightly around the diagonal (y=x), with a Pearson correlation coefficient of 0.873, indicating that there is a strong positive correlation between the joint image score output by the model and human subjective evaluation. The scatter distribution reveals that the aggregation of prediction points is better than that of the middle interval [2.5, 4.0] at both ends of the scoring interval [1.0, 2.5] (low coordination area) and [4.0, 5.0] (high coordination area). This phenomenon suggests that the model is more sensitive and accurate to the extreme "uncoordinated" and "very coordinated" scenes. However, there is still some uncertainty in the perception prediction of the middle fuzzy zone, and the prediction error in this region mainly comes from the large variance of manual scoring itself (the standard deviation of scoring of different subjects in the same scene is 0.43). In contrast, the Cross-Attention method shows a noticeable fan-shaped deviation from the diagonal, with systematic overestimation (by 0.32 on average) in the low-score region (<2.0) and underestimation (by 0.28 on average) in the high-score region (>4.5). It reflects the deficiency of standard attention mechanism in the ability of information screening between modes.

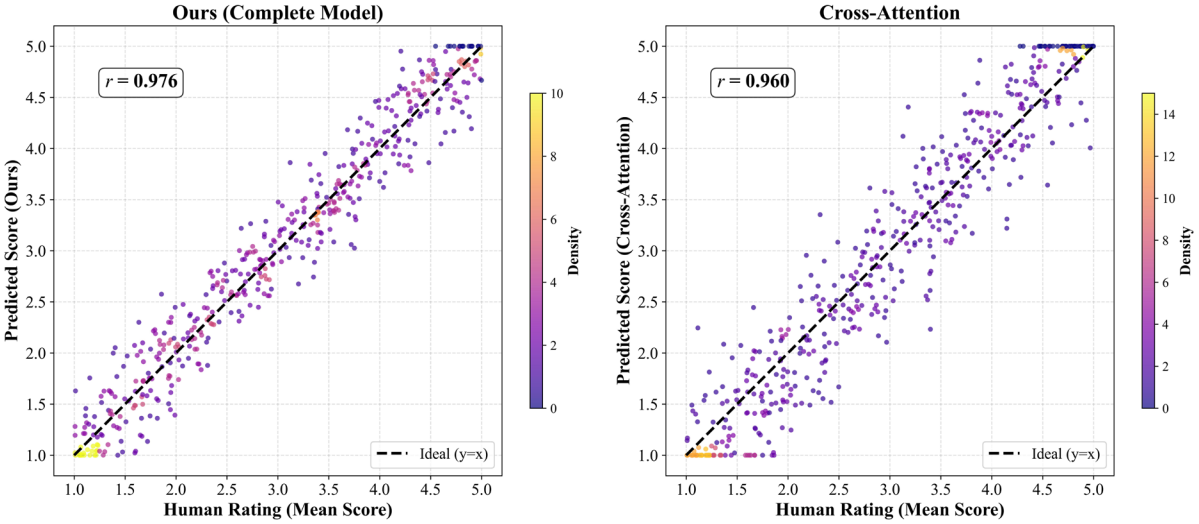

Figure 2. Scatter plot of perception score prediction: The left panel shows the full model and the right panel shows the Cross-attention method

We invited 30 subjects (16 male, 14 female) aged 23–45 (mean 31.2) to participate in the subjective perception validation. All had normal hearing and normal or corrected-to-normal vision, and possessed basic familiarity with Beijing's old city communities. We randomly selected 150 sampling points from the test set, along with their corresponding 10-s acoustic segments, and asked each subject to independently provide a joint impression score on the five-point scale for each audio-visual pair. To quantify consistency between model outputs and human judgments, we computed the intraclass correlation coefficient (ICC) using the two-way mixed model, absolute agreement definition, ICC(2,1):

$$\mathrm{ICC} = \frac{\mathrm{MS}_B - \mathrm{MS}_W}{\mathrm{MS}_B + (k-1)\mathrm{MS}_W + \frac{k}{n}(\mathrm{MS}_S - \mathrm{MS}_W)} \tag{26}$$

Where MS_B is the mean square between samples, MS_W is the mean square within samples, MS_S is the mean square between scorers, $k = 30$ is the number of subjects, and $n = 150$ is the number of samples. The human-model ICC was 0.824 (95% CI: [0.778, 0.861]), indicating excellent consistency (ICC > 0.75). We further divided the 30 subjects into two groups based on professional background in architecture or urban planning (12 professionals, 18 non-professionals). The ICC with the model was 0.851 for professionals and 0.809 for non-professionals, though the difference was not significant (p = 0.087). However, the slightly higher value for professionals suggests that the learned audio-visual association pattern is closer to the perceptual logic of trained observers. Table 3 summarizes the corresponding relationship between RMSE of model perception prediction and manual rating variance in each community type. It can be seen that RMSE and manual rating variance are positively correlated (Pearson r = 0.76), that is, in communities with large differences between subjects (such as mixed commercial and residential areas), the prediction accuracy of the model also decreases accordingly. This is in line with the law of expectations-subjective perceived uncertainty directly constitutes a ceiling on the predictive ability of the model.

Table 3. Comparison of RMSE and manual score variance of perception prediction for each community type

Type of community	Sample size	The perceived prediction RMSE	Standard deviation of manual score
Historic reserve	302	0.297	0.358
Mixed commercial and residential areas	318	0.351	0.427
Hutong micro Renewal Area	324	0.316	0.389
Traditional residential area	321	0.305	0.376

For significance testing, we used paired-sample t-tests to compare the mAP differences between the full model and each comparison method. For each test sample, we computed the AP values for the full model and the comparison method, then formed paired differences to test the null hypothesis $H_0: \mu_{\text{diff}} = 0$. against the alternative $H_1: \mu_{\text{diff}} > 0$ (full model better). The paired t-test versus Cross-Attention gave $t(1264) = 6.73,\ p < 0.001$; versus MFB, $t(1264) = 9.14,\ p < 0.001$; and versus the unimodal visual baseline, $t(1264) = 14.28,\ p < 0.001$. All p-values were far below 0.05, indicating highly significant improvements. We also performed one-way ANOVA on the mAP differences across the four community types, yielding $F(3,1261) = 8.52$, $p < 0.01$. Post-hoc Tukey HSD tests revealed a significant difference between the mixed commercial-residential area and the historical reserve (adjusted $p = 0.008$). However, differences among other community pairs did not reach significance, this statistically confirms that community spatial form has a modulating effect on model performance.

Finally, to provide intuitive qualitative insight, we selected eight typical scenes from the test set, those with the highest (>0.92) and lowest (<0.45) cosine similarity—for visual inspection. The high correlation scenes mainly present the following commonalities: hutong segments with a moderate sense of enclosure (aspect ratio ~0.6–1.2); visual content featuring traditional grey brick walls and greenery; acoustic background composed of bird songs and distant indistinct human voices; a flat noise spectrum without prominent peaks. This combination suggests a semantic resonance between the “softness” of the soundscape and the “traditional tranquility” conveyed by the visual scene. Low-similarity scenes fall into two categories. The first exhibits traditional visual features but contains continuous low-frequency mechanical vibrations in the soundscape (e.g., underground construction, outdoor air-conditioner clusters), which cause a sharp increase in psychoacoustic roughness (>2.3 asper) while the visual image remains “traditional” — leading to conflicting emotional information. The second category involves visually strong modern elements (e.g., neon signs) co-occurring with traditional street cries in the soundscape. These two modalities point to completely different temporal eras, canceling each other in the joint embedding space and resulting in low similarity.

5. CONCLUSION AND FUTURE WORK

Focusing on the cross-modal correlation between soundscape and visual imagery in Beijing’s old city communities, this paper has constructed a complete computational modeling framework covering data collection, feature extraction, model architecture, and experimental verification. Through synchronous acquisition with a mobile sensor array and DTW-based spatio-temporal alignment

preprocessing, we effectively solve the problem of accurately matching heterogeneous modal data in dense street environments. On this basis, we propose three core innovations—the cross-modal attention forgetting gate, the hierarchical contrastive loss, and the spatio-temporal adaptive modality weight regulator—which systematically enhance cross-modal correlation modeling from the perspectives of noise suppression, semantic alignment, and spatial adaptation. Experimental results demonstrate that the proposed model achieves a cross-modal retrieval mAP of 76.82%, outperforming the best comparison method by 8.68 percentage points; the RMSE of perception score prediction is as low as 0.318, and the intraclass correlation coefficient with human scores reaches 0.824, fully validating its effectiveness and practicality. In the quantitative feature response analysis, the model is most sensitive to the coupling between psychoacoustic roughness and visual building density, with a partial correlation coefficient exceeding 0.78 in the joint embedding space. This finding reveals a deep sensory coordination mechanism between auditory texture density and visual spatial enclosure in old city communities. As visual building density increases and space narrows, rapid modulation components in the soundscape tend to intensify simultaneously, together forming a multi-sensory expression of "crowdedness." This effect is particularly pronounced in the contrast between historical reserves and mixed commercial-residential areas.

Although this paper has made systematic progress in methodology and experimental validation, several limitations remain. First, data collection was conducted under relatively stable weather conditions (wind speed below Beaufort force 3, no precipitation); the model's robustness to multimodal data under extreme weather (e.g., microphone wind noise, acoustic attenuation, and degraded visual contrast due to heavy rain) has not been fully tested. Wind noise significantly raises low-frequency energy, interfering with accurate MFCC and roughness computation, while rain streaks occlude panoramic images and weaken visual feature extraction. The current model lacks specific countermeasures for such simultaneous modal degradation. Second, although our spatio-temporal adaptive weight regulator incorporates GPS spatial density and day-night factors, it does not account for long-term temporal variations (e.g., seasonal changes) or accidental transient events (e.g., construction impacts, festival loudspeakers), so the model's temporal generalization capability requires further expansion.

Future work will focus on the following directions. First, to address data degradation under extreme weather, we plan to propose a multi-modal data augmentation strategy using generative adversarial networks to simulate wind noise, rain, and fog effects, thereby enriching training samples and improving robustness in low SNR scenes. Second, the current model treats each sampling point independently, neglecting the spatial topology of old city communities—adjacent points exhibit natural smooth transitions and spatial dependencies in both soundscape and visual imagery. We will introduce graph neural networks, abstracting street networks as graphs with sampling points as nodes and edges defined by street connectivity and spatial distance. Graph convolution will aggregate neighborhood information, enabling the model to capture the propagation and attenuation patterns of audio-visual associations along street topologies, thus achieving a more macroscopic image understanding at the community level. Additionally, we are planning a data collection scheme combining long-term fixed monitoring nodes with mobile acquisition, to enable continuous tracking of the dynamic evolution of soundscape and visual imagery in old city communities, providing a more timely computational tool for multi-sensory quality assessment in urban renewal.

Abbreviations

DTW, Dynamic Time Warping;

GPS, Global Positioning System;

IMU, Inertial Measurement Unit;

MFCC, Mel-Frequency Cepstral Coefficients;

FFT, Fast Fourier Transform;

GRU, Gated Recurrent Unit;

Transformer, Transformer;

ResNet, Residual Network;

DeepLabV3+, DeepLab version 3 Plus;

CM-AFG, Cross-Modal Attention Forget Gate;

HCL, Hierarchical Contrastive Loss;

STAMW, Spatio-Temporal Adaptive Modality Weighting;

InfoNCE, Information Noise-Contrastive Estimation;

MLP, Multilayer Perceptron;

MFB, Multi-modal Factorized Bilinear Pooling;

mAP, mean Average Precision;

AP, Average Precision;

RMSE, Root Mean Square Error;

ICC, Intraclass Correlation Coefficient;

ANOVA, Analysis of Variance;

Tukey HSD, Tukey Honest Significant Difference;

SNR, Signal-to-Noise Ratio;

GNN, Graph Neural Network;

ReLU, Rectified Linear Unit;

AdamW, Adaptive Moment Estimation with Weight Decay;

GPU, Graphics Processing Unit;

CPU, Central Processing Unit;

EMA, Exponential Moving Average.

Supplementary Material

Not applicable.

Appendix

Not applicable.

Ethics approval and consent to participate.

This study did not involve human participants, animal subjects, or any data requiring ethical approval. Therefore, ethics approval and consent to participate are not applicable.

Acknowledgements

The authors would like to thank the editors of this journal and all the anonymous reviewers who provided valuable comments on this work.

Competing interests

The authors declare that they have no financial or personal relationships that may have inappropriately influenced them in writing this article.

Author contributions

All authors have read and agreed to the published version of the manuscript. The author's contributions are specified as follows: **S.Z.:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **Q.X.:** Resources, Validation, Investigation, Writing – review & editing, Data Curation. **L.M.:** Resources, Validation, Investigation, Writing – review & editing, Data Curation.

Funding information

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability

The data that support the findings of this study are available upon request from the corresponding authors, **L.M.**

Disclaimer

The views and opinions expressed in this article are those of the authors and are the product of professional research. It does not necessarily reflect the official policy or position of any affiliated institution, funder, agency, or that of the publisher. The authors are responsible for this article's results, findings, and content.

Declaration of AI and AI-assisted Technologies in the Writing Process

During the writing of this article, the author used DeepSeek for spelling and grammar checking. After using this tool, the author reviewed and edited the content as needed and assumes full responsibility for the final published content.

REFERENCES

[1] Wang, S., Zhang, J., Wang, F., & Dong, Y. (2023). How to achieve a balance between functional improvement and heritage conservation? A case study on the renewal of old Beijing city. *Sustainable Cities and Society*, *98*, 104790. https://doi.org/10.1016/j.scs.2023.104790

[2] Zhang, R., Martí Casanovas, M., Bosch González, M., & Sun, S. (2024). Revitalizing heritage: The role of urban morphology in creating public value in China's historic districts. *Land*, *13*(11), 1919. https://doi.org/10.3390/land13111919

[3] Amir, S., Sadoway, D., & Dommaraju, P. (2023). Taming the noise: Soundscape and livability in a technocratic city-state. *East Asian Science, Technology and Society: An International Journal*, *17*(1), 88-104. https://doi.org/10.1080/18752160.2021.1936749

[4] Gil-Sayas, S., Di Pierro, G., Tansini, A., Serra, S., Currò, D., Broatch, A., & Fontaras, G. (2024). Energy consumption of mobile air-conditioning systems in electrified vehicles under different ambient temperatures. *International Journal of Engine Research*, *25*(2), 293-304. https://dx.doi.org/10.1177/14680874231171303

[5] Bahgat, G., Al-Makhlasawy, R. M., Khairy, M., Nour, M., & Abdelfattah, A. (2026). Energy saving for air conditioning devices based on the amalgamation of intelligent models and occupancy detection. *Journal of Electrical Systems and Information Technology*, *13*(1), 78. https://doi.org/10.1186/s43067-026-00379-1

[6] Zhang, D., Ni, J., & Shi, X. (2024). Study of low-temperature energy consumption optimization of battery electric vehicle air conditioning systems considering blower efficiency. *Processes*, *12*(7), 1495. https://doi.org/10.3390/pr12071495

[7] Muhammad, K., Hussain, T., Ullah, H., Del Ser, J., Rezaei, M., Kumar, N., ... & De Albuquerque, V. H. C. (2022). Vision-based semantic segmentation in scene understanding for autonomous driving: Recent achievements, challenges, and outlooks. *IEEE Transactions on Intelligent Transportation Systems*, *23*(12), 22694-22715. https://doi.org/10.1109/TITS.2022.3207665

[8] Emek Soylu, B., Guzel, M. S., Bostanci, G. E., Ekinci, F., Asuroglu, T., & Acici, K. (2023). Deep-learning-based approaches for semantic segmentation of natural scene images: A review. *Electronics*, *12*(12), 2730. https://doi.org/10.3390/electronics12122730

[9] Liang, Y., Mitchell, A., Kang, J., & Aletta, F. (2026). A Review of Soundscape Datasets: Challenges

and Prospects for Multimodal Research. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2026.3659084

[10] Chen, M., Lin, Z., Song, X., Luo, Y., Duan, X., Li, S., & Xie, C. (2026). Theoretical Framework, Technical Evolution, and Future Prospects of Cross-Modal Mapping and Controllable Image Generation Under Multi-Source Heterogeneous Collaboration. *Sensors (Basel, Switzerland)*, *26*(10), 2972. https://doi.org/10.3390/s26102972

[11] Kothinti, S. R., & Elhilali, M. (2023). Are acoustics enough? Semantic effects on auditory salience in natural scenes. *Frontiers in Psychology*, *14*, 1276237. https://doi.org/10.3389/fpsyg.2023.1276237

[12] Sun, J., Deng, L., Afouras, T., Owens, A., & Davis, A. (2023). Eventfulness for interactive video alignment. *ACM Transactions on Graphics (TOG)*, *42*(4), 1-10. https://doi.org/10.1145/3592118

[13] Zhuang, Y., Kang, Y., Fei, T., Bian, M., & Du, Y. (2024). From hearing to seeing: Linking auditory and visual place perceptions with soundscape-to-image generative artificial intelligence. *Computers, Environment and Urban Systems*, *110*, 102122. https://doi.org/10.1016/j.compenvurbsys.2024.102122

[14] Chen, P., Huang, X., Fei, T., & Wang, S. (2026). Cross-Modal Urban Sensing: Evaluating Sound–Vision Alignment Across Street-Level and Aerial Imagery. *Transactions in GIS*, *30*(2), e70246. https://doi.org/10.1111/tgis.70246

[15] Zhang, Y., Wu, M., & Cai, X. (2025). A dynamic cross-modal learning framework for joint text-to-audio grounding and acoustic scene classification in smart city environments. *Digital Signal Processing*, *167*, 105444. https://doi.org/10.1016/j.dsp.2025.105444

[16] Wang, T., Li, F., Zhu, L., Li, J., Zhang, Z., & Shen, H. T. (2025). Cross-modal retrieval: a systematic review of methods and future directions. *Proceedings of the IEEE*, *112*(11), 1716-1754. https://doi.org/10.48550/arXiv.2308.14263

[17] Yang, Y., Wang, D., Hu, Y., & Liang, L. (2025, June). Research on Time Synchronization Technology of Non-safety DCS System Based on IEEE1588v2 Timing Protocol. In *International Conference on Nuclear Engineering* (pp. 163-176). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-95-2921-6_13

[18] Patoli, A. A., & Fortino, G. (2025). FPGA-based system implementation of IEEE 1588 precision time protocol: A review. *IEEE Sensors Journal*, *25*(11), 18624-18642. https://doi.org/10.1109/JSEN.2025.3557277

[19] Iturbe-Martin, Z., Martín-Garín, A., & Casado-Rezola, A. (2026). Soundscape-Informed Urban Planning and Architecture in Historic Centers: A Multi-Layer Method for Soundscape Characterization Applied to Bilbao Old Town. *Applied Sciences*, *16*(8), 3630. https://doi.org/10.3390/app16083630

[20] Versümer, S., Steffens, J., & Weinzierl, S. (2023). Day-to-day loudness assessments of indoor soundscapes: Exploring the impact of loudness indicators, person, and situation. *The Journal of the Acoustical Society of America*, *153*(5), 2956-2956. https://doi.org/10.1121/10.0019413

[21] Chen, Y., Lin, Y., Xu, R., & Vela, P. A. (2023, October). Wdiscood: Out-of-distribution detection via whitened linear discriminant analysis. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 5275-5284). IEEE. https://doi.org/10.48550/arXiv.2303.07543

[22] Gao, S., Yang, K., Shi, H., Wang, K., & Bai, J. (2022). Review on panoramic imaging and its

applications in scene understanding. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1-34. https://doi.org/10.1109/TIM.2022.3216675

[23] Taha, K. (2026). Generative AI for multimodal content: a survey with empirical and experimental evaluations. *Artificial Intelligence Review*, *59*(6), 140. https://doi.org/10.1007/s10462-026-11525-6

[24] Yan, T., Zhao, S., Hu, M., Wang, M., Zhang, X., Luo, Z., & Wang, M. (2024). HCL: A hierarchical contrastive learning framework for zero-shot relation extraction. *IEEE Transactions on Neural Networks and Learning Systems*, *36*(3), 5694-5705. https://doi.org/10.1109/TNNLS.2024.3379527

[25] Zhong, B., Wang, P., & Wang, X. (2024, August). Ts-HCL: hierarchical layer-wise contrastive learning for unsupervised domain adaptation on time-series. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data* (pp. 31-45). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-97-7238-4_3

[26] Xie, H., He, Y., Wu, X., & Lu, Y. (2022). Interplay between auditory and visual environments in historic districts: A big data approach based on social media. *Environment and Planning B: Urban Analytics and City Science*, *49*(4), 1245-1265. https://doi.org/10.1177/23998083211059838

[27] Yao, X. W., Zheng, J. M., Li, Q., Yang, K. H., Yao, Z. H., & Shang, Y. H. (2025, September). MSTDFN: Multi-modal Spatio-Temporal Dynamic Fusion Network for Traffic Flow Prediction. In *China Conference on Wireless Sensor Networks* (pp. 189-202). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-95-9615-7_12

[28] Zhao, T., Chen, G., Suraphee, S., Phoophiwfa, T., & Busababodhin, P. (2025). A hybrid TCN-XGBoost model for agricultural product market price forecasting. *PLoS One*, *20*(5), e0322496. https://doi.org/10.1371/journal.pone.0322496

[29] Zhao, T., Chen, G., Pang, C., & Busababodhin, P. (2025). Application and performance optimization of SLHS-TCN-XGBoost model in power demand forecasting. *Comput. Model. Eng. Sci*, *143*(3), 2883-2917. https://doi.org/10.32604/cmes.2025.066442

[30] Zhao, T., Chen, G., Pang, C., Seenoi, P., Papukdee, N., & Busababodhin, P. (2025). Time-lapse earthquake difference prediction based on physics-informed long short-term memory coupled with interpretability boosting. *Journal of Seismic Exploration*, *34*(3), 25. https://doi.org/10.36922/JSE025310049

[31] Zhao, T., Chen, G., Pang, C., Li, L., & Busababodhin, P. (2026). Forecasting global agricultural trade imbalances using a hybrid deep learning and gradient boosting framework. *Discover Computing*, *29*(1), 443. https://doi.org/10.1007/s10791-026-10367-8

APA Citation

Shen, Z., Qi, X., & Liu, M. (2026). Computational Modeling of Soundscape and Visual Imagery in Beijing Old City Communities Based on Multi-modal Data. *Journal of Digital Frontier, 1*(1), 61-82. **https://doi.org/10.67541/JDF2603**