"Virtual and Real Coexisting": Zither Composition Practice and Aesthetic Reconstruction Through AI and VR Integration

Liulingzi Xiao[1], Wenzhen Xiong[2], Haibo Zhang[3]
[1]College of Humanities and Arts, Changde College, Changde 415000, Hunan, China
[2]Department of General Education, Changde Vocational Technical College of Science and Technology, Changde 415000, Hunan, China
[3]Art Academy, Hunan College for Preschool Education, Changde 415000, Hunan, China

Corresponding author: Liulingzi Xiao (e-mail: 13521191075@163.com)

ABSTRACT This study presents a multi-modal signal processing framework for automatic guzheng composition and VR-based performance simulation. The main melody is extracted using a similarity matrix, while LSTM-DQN networks, combined with quantized reward mechanisms based on musical theory and playing techniques, generate harmonic sequences. Multi-channel sensory data—including audio, tactile feedback, hand gesture, and line-of-sight tracking—are processed and integrated to construct an interactive VR performance environment, forming a closed-loop feedback system analogous to multi-channel signal transmission and control in engineering. Experimental results demonstrate that the LSTM-DQN model achieves an average note prediction accuracy of 67.6% (σ=2.34%) and maintains high rhythmic fidelity, while user evaluations indicate satisfactory aesthetic quality (average score 3.04/5). The framework enables rapid transformation from compositional ideas to virtual performance outputs, highlighting the effectiveness of multisensor signal integration, temporal sequence learning, and real-time feedback in complex systems. This approach provides an engineering-oriented perspective for integrating AI-driven generative models with interactive feedback, supporting future applications in intelligent multi-channel signal processing and virtual performance control systems.

INDEX TERMS LSTM-DQN, multi-channel signal processing, VR feedback, generative music systems

I. INTRODUCTION

THE "zheng" can be traced back to the Qin Dynasty, and reached unprecedented prosperity and splendor in the Tang Dynasty, and began to be widely circulated in the folk, becoming an important medium for people's recreation and leisure, and expressing their emotions [1]. In the history of more than two thousand years, "zheng" has not only accumulated a wealth of performance skills and repertoire, but also incorporated a deep cultural heritage and national sentiment, which has been constantly precipitated, developed, and finally formed a deep cultural heritage and unique national temperament [2,3]. It is not only a treasure in the treasure house of Chinese music, but also an important carrier to show the unique temperament and aesthetic interest of the Chinese nation. Since the twentieth century, digital intelligence technology has driven artists to transform traditional crafts by integrating advanced science with Chinese heritage, evolving musical performance techniques to create national works that define the modern era [4,5]. The development of artificial intelligence (AI) technology can be traced back to the 1940s and 1950s, and with the significant increase in computing power and data volume, AI has made rapid progress in the past decades [6]. The core technologies of AI include machine learning and deep learning. Machine learning parses large amounts of data and recognizes patterns and regularities to improve task execution through algorithms that can be categorized as supervised, unsupervised, and reinforcement learning [7]. Deep learning, on the other hand, is able to handle highly complex data by constructing and training multi-layer neural networks, such as image and speech recognition, and the application of convolutional neural networks in image recognition and recurrent neural networks in natural language processing demonstrates the powerful ability of deep learning [8-10]. In recent years, some scholars have studied the Chinese guzheng in order to explore its cultural value. Yang analyzed the creation background and expression methods of the guzheng work "Night Mooring on the Maple Bridge", which is also an ancient poem during the Tang Dynasty. This art work combining the poem and the guzheng includes a variety of performance techniques, which succeeded in revealing the emotions that the original

poet wanted to express [11]. Chen et al. established a multifactor regression analysis model in order to study the overall influence of guzheng music playing style on Chinese traditional music culture, revealing the influence of the characteristics of guzheng music, such as historical records, creations, and regional influences, on Chinese traditional music culture, and providing effective information for the promotion of Chinese traditional music [12]. Ning et al. studied the historical evolution of Chinese guzheng music creation through interview and observation methods, and successfully explored the characteristics and expressions of guzheng music creation, where composers and performers jointly realized guzheng music creation, and the diversity of performance techniques was the key to its artistic expression [13]. Li et al. conducted a systematic study on the origin and development of the guzheng, including the origin of the word "zheng", the meaning of the proposition, and the historical and cultural connotations, based on which they put forward a proposal for the protection and inheritance of the guzheng's music and culture, so that the varied and rich culture of the guzheng can be adapted to the development of the modern society [14]. Zhang, Y et al. designed a new automatic music generation network, the network combines layer normalized recursive skip-connection model, Transformer-XL model, and multi-threaded attention mechanism, which attenuates the problem of gradient vanishing and at the same time increases the ability of music feature capture, and the test results also verified the feasibility of the proposed model in music generation [15]. This paper extracts the main melody from guzheng music using a similarity matrix and converts it into MIDI format, utilizing an LSTM network to generate notes. This algorithmic approach ensures the musical sequence maintains a rhythmic consistency comparable to the precision of digitized fabric weaving. Based on Deep Reinforcement Learning DQN algorithm, based on the theory and technique of guzheng music, we design the reward quantization mechanism of guzheng music, generate guzheng features, and input the melody and features of guzheng music into the virtual scene. The virtual scene of guzheng playing is constructed through VR technology, and the multi-channel interaction function is proposed, and the interactions of auditory, tactile, playing gesture, and line-of-sight tracking are designed respectively to realize the VR-based guzheng playing scene. Construct the zheng music dataset and measure the composing ability of the network proposed in this paper through simulation experiments. Construct the backbone sound framework and perform cross-validation test and amplification test on zheng tunes to evaluate the quality of zheng tune generation. The generated zheng tunes are evaluated and scored by the public to assess the aesthetic and artistic value of the generated zheng tunes.

II. ZITHER COMPOSITION BASED ON AI TECHNOLOGY

A. PRE-PROCESSING OF GUZHENG MUSIC DATA

1) Introduction and Extraction of Choruses

In order to obtain the main theme of the music, starting from its expression, this paper uses a similarity matrix to extract the main theme due to its self-replicating nature. In MIDI format, if the music is scanned at 10 frames per second, i.e., 1 frame per 0.1 second. Each frame of music can be described by a vector $v = [x_1, x_2, \ldots, x_n]$, where n denotes the number of notes contained inside a frame of music, and a piece of music consists of multiple frames of music, so it can be described by a matrix $X = [v_1, v_2, \ldots, v_N]^T$, where it is assumed that each piece of music has N frames. It follows that the similarity matrix of a piece of music is the $(N-1) \times (N-1)$ matrix, and the distance between two pieces of music can be calculated using the Euclidean distance:

$$S_{i,j} = \sqrt{\left(v_i - v_{i+1}\right)^2 + \left(v_j - v_{j+1}\right)^2} \tag{1}$$

where $i, j = 1, 2, \ldots, N-1$.

The computation of the similarity matrix is good for finding repeated segments with small variations, which can be retrieved using the formula defined as follows:

$$S = 1 - \frac{\left\| X_i - X_{i+1} \right\|}{\sqrt{12}} \tag{2}$$

Where 12 is the twelve mean law, X_i and X_{i+1} are the matrices corresponding to the two pieces of music. If the calculation results in a high value of S, it means that the two pieces of music are similar, and the similar segments will be used in Python for the extraction of the chorus part of that part, which will be stored in the trained guzheng database.

B. DESIGN AND TRAINING OF LSTM NETWORKS

In this paper, LSTM network will be used for the generation of musical notes for guzheng music. In the following, the computation of LSTM cells will be described in detail. For a given moment input to the LSTM, three different input modalities will be included at the same time: the data input (labeled vector) x_t of this cell, and the outputs h_{t-1} and c_{t-1} of the previous cell. Where h_{t-1} is the previous cell output and c_{t-1} denotes the extent to which the previous cell output should be retained. Figure 1 shows a schematic of a cell of the LSTM, which consists of three gates: the oblivion gate (blue box), the input gate (pink box), and the output gate (yellow box). LSTM Forgetting Gate: Used to control how much information will be "forgotten" from the previous cell, which is calculated as follows:

$$f_t = \delta\left(w_f \times [h_{t-1}, x_t] + b_f\right) \tag{3}$$

Where h_{t-1} is the output of the previous cell, x_t is the input to this cell, w_f is the weights, b_f is the bias parameter, and δ is the Sigmoid function. f_t multiplied by c_{t-1} determines how much information from the previous cell is kept for

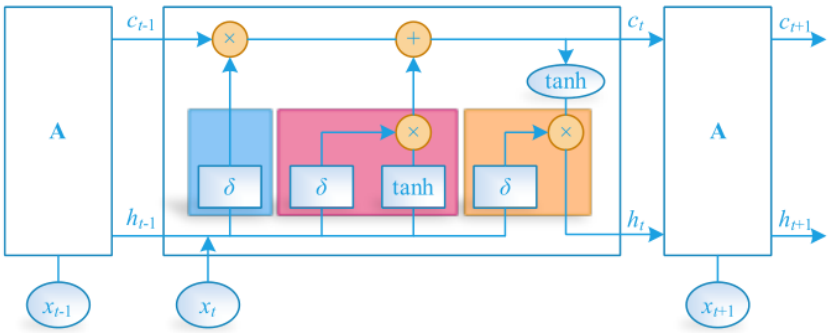


FIGURE 1. One LSTM unit

this cell. LSTM Input Gate: It is used to determine the input situation and it consists of two parts:

$$\begin{aligned} i_t &= \delta\left(w_i \times [h_{t-1}, x_t] + b_i\right), \\ c_t &= \tanh\left(w_c \times [h_{t-1}, x_t] + b_c\right). \end{aligned} \tag{4}$$

Where w_f and w_i are weights and b_i and b_c are bias parameters. LSTM output gate: used to determine the information output of this unit. Just like the previous unit, it gives two outputs: h_t is the actual output and c_t indicates how much of h_t should be kept. The output h_t is computed as follows:

$$o_t = \delta\left(w_o \times [h_{t-1}, x_t] + b_o\right). \tag{5}$$

$$h_t = o_t \times \tanh(c_t). \tag{6}$$

where w_o is the weight and b_o is the bias parameter. The control c_t is calculated as follows:

$$c_t = f_t \times C_{t-1} + i_t \times c_t. \tag{7}$$

It is worth noting that when $f_t \to 1$ and $i_t \to 0$, $c_t \approx c_{t-1}$, all the information from the previous cell will be preserved. Conversely, when $f_t \to 0$ and $i_t \to 1$, $c_t \approx c_t$, all the information of the previous cell will not be preserved. For a given dataset $\{h_{j,t}, j = 1, 2, \ldots, M, t = 1, 2, \ldots, N\}$, the probabilities are first calculated:

$$\begin{aligned} P(h_{j,t}) &= \frac{e^{h_{j,t}}}{\sum_{j=1}^{M} e^{h_{j,t}}}, \\ j &= 1, 2, \ldots, M, \quad t = 1, 2, \ldots, N \end{aligned} \tag{8}$$

Then, the note with the highest probability is selected as the generated note for the LSTM:

$$y_t = \arg\max_j P(h_{j,t}), \qquad t = 1, 2, \ldots, N \tag{9}$$

The final result of the training convergence of the LSTM is based on the minimization of the cross-entropy loss function:

$$L_1 = -\frac{1}{M}\sum_{j=1}^{M}\left(Y_j \log(Y_j) + (1 - Y_j)\log(1 - Y_j)\right) \tag{10}$$

$$Y_j = \frac{1}{N}\sum_{t=1}^{N}(x_{j,t} - y_{j,t})^2 \tag{11}$$

C. GUZHENG FEATURE GENERATION BASED ON DEEP REINFORCEMENT LEARNING DQN ALGORITHM

1) Quantification of Reward and Punishment Mechanisms for Guzheng Music

In this chapter, the maximum probability note generated by the LSTM network at the current moment is defined as the selected action y_t (9), where $y_t \in A$, A is the set of all actions. Packing all notes generated before moment t is considered as the previous state s_{t-1}, where:

$$(y_1, y_2, \ldots, y_{t-1}) \to s_{t-1}, \qquad s \in S. \tag{12}$$

The DQN network will combine the note a_t selected at the current t moment and the previous state s_{t-1} to computationally generate the next moment's state s_t. The DNN network in the DQN network will store s_{t-1}, y_t, and s_t, which will be used to determine the current action y_t is scored and provides storage for future randomized actions. Meanwhile, r_t^i is the reward and punishment scores obtained under different rules i.

2) Reward Mechanisms for Guzheng Music Theory Rules

(1) The reward and punishment mechanism of octave: Octave refers to the relationship between two tones containing the same name in the neighboring tone group, which is called octave. Most of the music styles of guzheng belong to classical music, unlike metal music, pop music and other styles, guzheng music does not need to produce strong changes between notes. In order to make the music sound smooth and rhythmically soothing, composers usually keep the arrangement of two neighboring notes within an octave, i.e., the interval difference needs to be less than an octave. In this paper, the treatment of octaves is similar to the extraction of musical choruses, and the twelve-mean-law pitch notation is used for the marking process, i.e., the intervals are between the numbers 12. According to the definition, between two neighboring notes, the intervals are within 12 to construct the reward and punishment strategy:

$$r_t^1(s_t, y_t, t) = \begin{cases} 0.7 & \text{When } |y_t - y_{t-1}| \le 12 \\ -0.8 & \text{When } |y_t - y_{t-1}| \ge 12 \end{cases} \tag{13}$$

(2) The reward and punishment mechanism of scales: Scales are octave scales composed of neighboring tones, usually pentatonic and heptatonic. The guzheng, as a traditional Chinese folk instrument, has a standard 21-string structure arranged in pentatonic scale (CDEGA), which in most of the playing conditions presents the simple scale 1, 2, 3, 5, 6 (do, re, mi, so, la), and in order to play the seven-tone scale (CDEFGAB), "fa(#F)" and "si(B)" in the seven-tone scale (CDEFGAB), it is usually obtained by pressing the strings with the left hand on the corresponding 'mi' (3) and 'la' (6) strings within the 21-string layout. According to the unique characteristics of the guzheng, the number of "fa(F)" and "si(B)" tones is defined to reduce their occurrence, and the reward and punishment strategies are constructed to penalize these non-pentatonic notes:

$$R_{\text{scale}} = \begin{cases} -0.8, & \text{if note is F or B,} \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

(3) Reward and punishment mechanism of base modes: The base modes are C major, D major, G major, etc. Each music has its base mode, and each mode has its core tone, such as D major. Each music has its base tuning, and each tuning has its core tone, such as the music in D major, whose dominant tone is D. In order to make the generated guzheng music more stable, the last generated tone of the automatic composition will be used as the dominant tone of that music. Therefore, according to the definition of the base tuning, the generated music of guzheng music needs to be set as the dominant tone of D major and as the ending tone of the guzheng music to construct the reward and punishment strategies:

$$r_t^3(s_t, y_t, t) = \begin{cases} 0.5 & \text{When } y_{t=max} \in \{\text{D}\} \\ -0.5 & \text{When } y_{t=max} \notin \{\text{D}\} \end{cases} \tag{15}$$

3) Reward Mechanisms for Guzheng Technique Rules

The technique rewarding mechanism is designed to add fingering techniques to the guzheng, as a way to capture the characteristics of guzheng music. (1) Arpeggio reward and punishment strategy: usually this is a technique consisting of three notes (the simplest arpeggio), each note separated by 1, and many arpeggios in guzheng music are located at the end of the sentence or at the end of the sentence, so according to the positional impact of the appearance of arpeggios, arpeggios appear to be the least frequency compared to the other two techniques. Therefore, assuming that at moment t, the input state is $s_t = \{A_t^1, A_t^2, A_t^3\}$, where A_t^1, A_t^2, A_t^3 denotes the notes produced at moment t and the positional arrangement conforms to the arpeggio arrangement order, the reward strategy is defined as follows:

$$r_t^4(s_t, y_t, t) = \begin{cases} 0.2, & \text{When } y_{t_{\max}} \in \{A_t^1, A_t^2, A_t^3\}, \\ -0.2, & \text{Other situations} \end{cases} \tag{16}$$

(2) Finger shaking (Yao Zhi) reward strategy: When a sequence of rapid identical notes appears, it is defined as the "Yao Zhi" technique. This technique is specifically designed to create a sustained, continuous sound through high-frequency repetition. To accurately reflect this musical purpose, the reward factor is applied to sustained repetitions that allow the zither to produce long-tone melodic lines, rather than penalizing sequences exceeding 6 notes. This ensures the model can simulate the characteristic tremolo effect essential for expressive Guzheng performance. The reward policy for finger shaking, r_t^5, is optimized to encourage rhythmic density and velocity consistency in repeated note sequences, enabling the generation of seamless long tones. The model recognizes these sequences as a single expressive unit rather than isolated repetition., when the finger shaking situation is met, the batch of notes, according to the MIDI file's Time Format Type: Precision, which indicates that the number of tick time units that each beat (quarter-note) is split into is halved, making the segment sound more like a finger shaking technique:

$$r_t^5(s_t, y_t, t) = \begin{cases} 0.3, & \text{When } y_t = y_{t-1} = y_{t-2} = y_{t-3}, \\ 0, & \text{When } y_t \neq y_{t-1}, \\ -1, & \text{When } y_t = y_{t-1} = y_{t-2} = y_{t-3} \\ & \quad = y_{t-4} = y_{t-5}. \end{cases} \tag{17}$$

(3) Major and Minor Handicap Reward and Punishment Strategies: Both major and minor handfuls contain two notes. The major clusters usually appear as odd-numbered strong beats, and the minor clusters usually appear in even-numbered weak beat positions. These two techniques are widely used in the more junior guzheng examinations, and are therefore expected to be heavily featured in the final generated music, which is rewarded by the technique:

$$r_t^6(s_t, y_t, t) = \begin{cases} 0.6, \text{if } |y_t - y_{t-1}| = 12 \\ \quad \text{and } t \mid 2 = 1, \\ 0.4, \text{if } |y_t - y_{t-1}| = 1 \\ \quad \text{and } t \mid 2 = 0. \end{cases} \tag{18}$$

This completes the generation of guzheng features based on deep reinforcement learning DQN algorithm.

III. VIRTUAL SCENE CONSTRUCTION OF GUZHENG PERFORMANCE UNDER VR TECHNOLOGY

A. VR GUZHENG PLAYING MULTI-CHANNEL INTERACTION

1) Zither Sound Feedback in the Auditory Channel

Based on the informative mastery and understanding of guzheng sound, this paper can select each piece of music and design the sound feedback of its zheng music. After the selection of tracks is completed, it is necessary to collect the sound of each zheng song, which needs to be completed in a professional recording studio, and the prior collection of each zheng song is also conducive to the adjustment of the 3D gesture guidance animation in the later stage. After the tuning process of Adobe Audition, the time seconds of each piece of music will be recorded meticulously, in order

to determine the start and end time of each fingering in the subtractive scores, and at the same time, the resting beats will be recorded, and with the help of the visual function of the audio track in Adobe Audition, it will be convenient to review the sound material after the processing of the method, and the sound material will be equipped as the basis of the gesture-guided animation. After this method, the sound material will be able to be used as the basis of gesture guidance animation, and it is the synchronization of sound and picture that can make the feedback of the sound channel more accurate and natural. After the sound of each zheng piece is captured, the sound material of each piece will be integrated into the sound output port of this paper, so that the user can hear the accurate timbre and rhythm of the zheng piece in time when he/she is using the module of each piece in the model, so as to deepen his/her perception of the timbre of the guzheng, and to get the supply of the content of the auditory channel. In the playing module of the model, the experiencer does not have to worry about the wrong fingering disrupting the normal feedback of the zheng sound, and the continuous playing of the zheng sound is also conducive to the completeness of the mood during the playing experience.

2) Tactile Perception Based on Physical Objects of the Koto

The loading of the physical guzheng in the model also needs to be realized with the help of Vive Tracker. The model needs to realize the position update and positioning of the physical guzheng, in order to correspond to the scene expansion in the later stage of the model, then the guzheng loaded with Vive Tracker can satisfy the realization of any spatial movement, and when the experiencer shifts the physical guzheng, the zither in the virtual scene can also complete the movement, which can provide the experiencer the freedom of action when stroking the zither and singing.

3) Gesture Recognition Interactions during Playing

The gesture recognition channel needs two links of design support, first of all, it needs to adjust the gesture guiding animation, which requires the recording of the timbre of each zheng song at the same time, the use of multi-angle video camera to complete the accurate grasp of the fingering of both hands when playing each zheng song, and the production of hand guiding animation to adjust the careful viewing of the comparison, so that the guiding animation of the hand gesture is accurate and error-free. When the final model is presented, attention should also be paid to the material of the guiding gesture animation, which needs to be presented with a translucent texture, so that it is easy to distinguish between the Leap Motion recognized hand and the guiding animation hand.

4) Selection Interactions Based on Line-of-Sight Tracking

Usually the development of a set of VR interactive model is required to provide point-and-click functionality with the support of VR handle, but when the model has multi-scene, multi-people experience needs, adding VR handle settings will increase the uncertainty of the model experience. For example, VR handle power loss, or artificial damage lost, multi-people use failed to timely disinfection, etc., the occurrence of these situations will greatly affect the smoothness of the experience of the VR interactive model. In view of this, this paper chooses line-of-sight tracking as the interaction realization goal in the click interaction means, in order to reduce the overloading of hardware devices and make the overall stability of the model stronger.

B. REALIZATION OF VR-BASED GUZHENG PLAYING SCENE

1) Gesture Tracking Recognition

According to the characteristics of the two modes of gesture input, the work in this paper adopts the method of combining the visual color feature recognition tracking with the data glove gesture recognition in two ways, to realize the real-time tracking and positioning based on the visual color feature, and to maintain real-time state gesture tracking according to the input of the data glove, and finally to realize the playing of the virtual guzheng according to the gesture of the data glove on the basis of this work. The data glove adopts the CAS-GLOVE data glove developed by the Institute of Automation of the Chinese Academy of Sciences, with 27 sensors. The data glove turns the sensor data into data that can be read by the computer. Vision-based gesture input is to use a single camera to capture the gesture object, and then use computer vision technology to analyze the captured image to extract the gesture image features, so as to realize the gesture input. Therefore, the gesture input in this paper adopts the data glove and monocular vision input in two ways, from the monocular vision gesture recognition tracking, it can be divided into hand localization and finger tracking. According to the hand model, we can know that the hand model space is very complex, the hand has 27 degrees of freedom, so the monocular vision in the hand recognition tracking inevitably meets the obstruction and difficult to effectively realize the difficulties, so now the finger state tracking and state recognition in the monocular vision is difficult to effectively realize, but also particularly complex, so this paper combined with the actual situation of data gloves on finger state tracking reflection, and the use of monocular vision to track the finger state, and the use of monocular vision to track the finger state. Therefore, this paper adopts the data glove to track the finger state in the practical situation, and uses monocular vision to localize the hand, so considering the superiority of the vision-based and data glove-based methods and the complementarity between them, combining the two organically together, thus playing the effect of "taking the strengths and complementing the weaknesses". In this way, the difficulty of monocular vision in handling finger tracking is avoided, and the combination of data glove processing effectively realizes the combination of the two methods to realize the recognition and tracking of hand gestures.

2) Video Image Acquisition and Image Processing

(1) Image plane tracking to identify the positioning of the center of gravity point calculation. The image is segmented to form a binarized image, according to the needs of this paper the color value of the target region as 255, while the background region has a color value of 0, the center point of which is the center of mass point of the target, which is calculated as follows. After segmenting the hand region, the pixel points in the region are binarized:

$$S(x',y') = \begin{cases} 1, & \text{If } (x',y') \text{ is the hand,} \\ 0, & \text{Reverse.} \end{cases} \tag{19}$$

This in turn determines the position of the hand's center of mass $(x'_{\text{hand}}, y'_{\text{hand}})$:

$$x'_{\text{hand}} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} S(i,j) \times i}{\sum_{i=1}^{N}\sum_{j=1}^{M} S(i,j)} \tag{20}$$

$$y'_{\text{hand}} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} S(i,j) \times j}{\sum_{i=1}^{N}\sum_{j=1}^{M} S(i,j)} \tag{21}$$

Where i and j are the number of row pixels and the number of column pixels in the identified manpower region, respectively.

(2) Distance reflection under monocular vision. Assuming that d is the distance of the target image from the camera, h is the side height of the target under vision, H is the target image height, D is the distance of the actual scene plane from the center of the camera, and f is the focal distance of the camera, so the distance d of the target image from the camera can be reflected by the following formula, target distance from the camera: focal distance = target side height: target image height on, as shown in Eq. (22) shown:

$$d : f = h : H \tag{22}$$

Based on the above equation one can calculate $d = f * h / H$.

Mapping visual image information to virtual scenes. The kinematic relationships of the various parts of the virtual hand joints can be illustrated by mathematical formulas. Since the mathematical models of the movements of the thumb, index finger, middle finger, ring finger, and pinky finger are similar, in this paper, only the wrist joint is illustrated. The base coordinates defined at the wrist joint are (X_0, Y_0, Z_0), with the X_0 axis pointing to the metacarpal bone along the forearm, which is the counterpart of the index finger. The Z_0 axis is parallel to the axis of rotation of the wrist joint, and the Y_0 axis is defined according to the right-hand rule. Define the wrist coordinate system WR on the wrist bone as (X_{WR}, Y_{WR}, Z_{WR}), Φ_{m} is the angle of rotation of the wrist joint around the X_{WR} axis. When $\Phi_{\text{m}} = 0°$, the palm and forearm are in a straight line and the base coordinate system coincides with the wrist coordinate system. From this, the chi-square transformation matrix from the wrist coordinates WR to the base coordinates can be derived:

$$^{0}_{WR}R = \begin{bmatrix} \cos(\phi_{WR}) & -\sin(\phi_{WR}) & 0 & 0 \\ \sin(\phi_{WR}) & \cos(\phi_{WR}) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{23}$$

The display plane is a plane parallel to the XOY plane of the virtual 3D coordinate system and perpendicular to the Z-axis, and the center of the display plane is on the Z-axis. Then the mapping of the image information to the virtual 3D coordinate system is converted to the mapping of the image information to the display plane, which constitutes the mapping of the 2D image coordinate points to the display plane points, and the conversion is realized according to the matrix transformation formula (24):

$$\begin{bmatrix} x^* \\ y^* \end{bmatrix} = k \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} 260 \\ 380 \end{bmatrix}. \tag{24}$$

The constant $k = 360/1.4$ was obtained from experiments and calculations.

IV. ZITHER COMPOSITION PRACTICE AND AESTHETIC EVALUATION ANALYSIS

A. EXPERIMENTAL ANALYSIS OF ZITHER COMPOSITION PRACTICE

The guzheng melody and guzheng features extracted by AI technology are inputted into the virtual scene of VR performance to achieve the purpose of playing automatically generated zheng tunes in the virtual scene.

1) Data Sets

The automatic composition network proposed in this paper was performed on a dataset of Guzheng compositions, which has been transcribed into a multi-layered texture comprising high, middle, and low registers. This structure reflects the polyphonic characteristics of Guzheng performance, where the right hand typically handles the high-register melody while the left hand provides middle-to-low register accompaniment and harmonic support. The data was processed to have a total of 350 segments after removing the choral parts with instrumental parts and the choral parts containing both notes, and then the data set was extended by transposing all the scores within the initial range, so that the final whole data set contained a total of 2,515 segments, of which 80% was used as the training set and the remaining 20% as the validation set. For each fragment, it consists of four voices together, which contain the number of note types 55, 57, 57, and 76, respectively.

2) Experimental Setup

In this paper, two kinds of experiments are conducted: (1) Experiment 1 The first is a numerical form of harmonic

overdubbing experiment, which is conducted for each of the four voices. Specifically, when the experiment is conducted for a particular voice, the remaining three voices are kept to use the real congregational hymn data while the notes of the current voice are all set to 0. Based on the supplementation of the congregational hymn metadata, the prediction is made after several iterations of the voice, and the number of iterations is set as [1200, 1300, 1400, 1500, 1800, 2200, 2300], using the first 100 pieces of music in the validation set as the test set, the predicted results were compared and calculated with the original piece of music according to the notes at the corresponding moments, and the average accuracy of these 100 pieces was taken as the final result, and compared with the reproduced results of LSTM-DQN. (2) Experiment two The second is a manual evaluation experiment, in which the test sample set is constructed in three ways: randomly sampling original works by acclaimed Guzheng masters (such as "Fighting the Typhoon" and "High Mountain and Flowing Water"), utilizing random sequence generation, and reproducing the LSTM-DQN to obtain 10 compositions each respectively, and disrupting the order of them to form a test set containing 30 samples, in which all the compositions are set to have a sequence length of 256 (16 bars), and the random sequence generation The experiment was set up with 10 sampling times [2200, 2700, 3700, 5000, 6600, 8000, 10000, 12000, 15000, 18000]. There were 15 subjects in the manual assessment experiment, which contained 5 musicians in the direction of guzheng, 5 musicians in the direction of non-guzheng and 5 music lovers. All the testers listened to all 30 pieces in order and judged each piece to be a human-composed masterpiece or computer-generated, and the testers could listen to a piece repeatedly before making a final decision on it, but could not modify their choices once they made a selection, and the selection results of all the testers were counted to measure the generative performance of the network proposed in this paper. This study was approved by the Ethics Committee of Changde College, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

3) Experimental Results

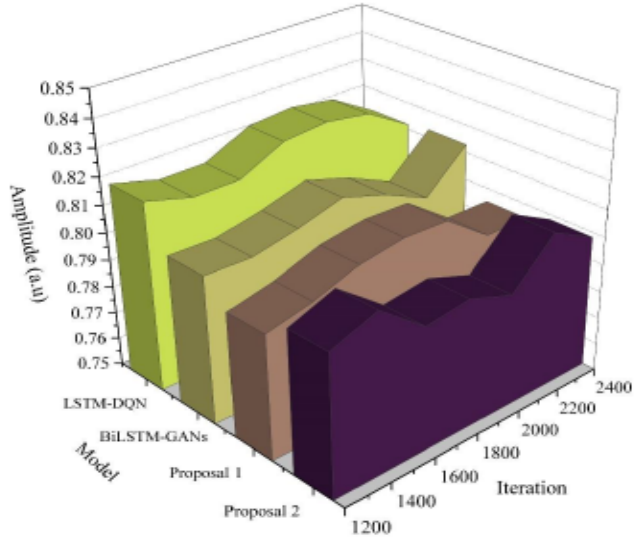
(1) Experiment 1 Fig. 2 shows the comparison of model accuracy, Fig. (a) shows soprano, Fig. (b) shows bass soprano, Fig. (c) shows tenor, and Fig. (d) shows bass baritone. The prediction accuracy of the model LSTM-DQN proposed in this paper for the soprano and baritone voices is always higher than the effect of the other two models and BiLSTM-GANs, with an average prediction accuracy of 0.818 and 0.748, respectively. and in the case of the soprano and the tenor, the accuracy curves of this paper's model and the BiLSTM-GANs have a crossover, which indicates that for the soprano and the tenor voices This paper's model also has some advantages, in which the advantage of the alto voice is obvious when the number of iterations is less, while the opposite is true for the tenor voice, which shows better performance when the number of iterations is more. Further, observing the accuracy curves of LSTM-DQN and proposal2, it can be found that, except for the tenor voice when the two curves have obvious crossings, proposal2 is basically lower than LSTM-DQN for the rest of the voice cases, and this paper's model will obtain more and better information about the piece by filtering the left and right features of the guzheng, and more accurately carry out the prediction without affecting the middle features. Prediction. Observing the accuracy curves of LSTM-DQN and proposal 2, it can be found that the curves of proposal 2 are always lower than that of LSTM-DQN for all the voice parts, which indicates that the addition of LSTM in the discriminative model obtains better temporal features and meets the needs of the discriminative model for feature extraction. The comparison of model accuracy is shown in Figure 2:

(2) Experiment II

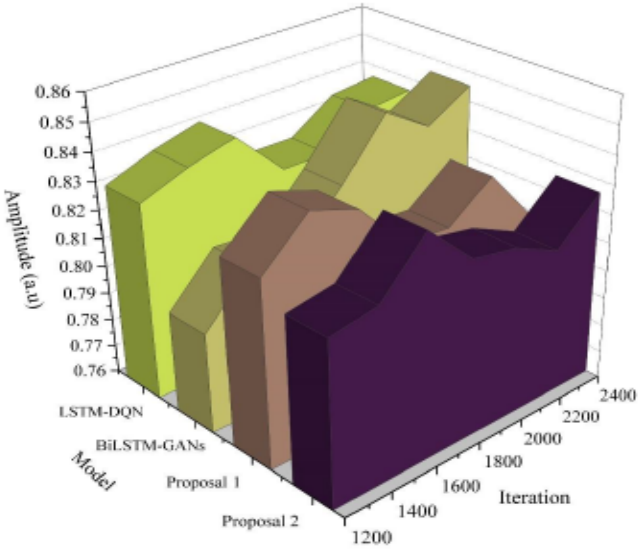
The generative model of LSTM-DQNguzheng arrangeme nt was used to generate multiple 16-bar-long pieces using multiple iterative sampling, which were manually evaluate d by disrupting the order with authentic human-compose d Guzheng classics and BiLSTM-GANs reproduction re sults. Figure 3 shows the results of manual evaluation, wh ere "BiLSTM-GANs" represents the music pieces obtaine d through reproduction, "Original" represents classical Guz heng repertoire, and "LSTM-DQN" represents the music pi eces generated from random sequences using the LSTM-DQN network. The scores generated using the model pro posed in this paper are lower than the BiLSTM-GANs res ults under the category of non-guzheng musicians,and high er under the rest of the categories. The vote rate is 5 1.748% and 51.654%, respectively, and, the music generate d by the randomized sequence has a lower vote rate under t he non-GuZheng musician category, and the rest of the cat egories and the combined vote rate can be close to 45%, which basically achieves a better auditory effect.

4) Backbone Frame Testing

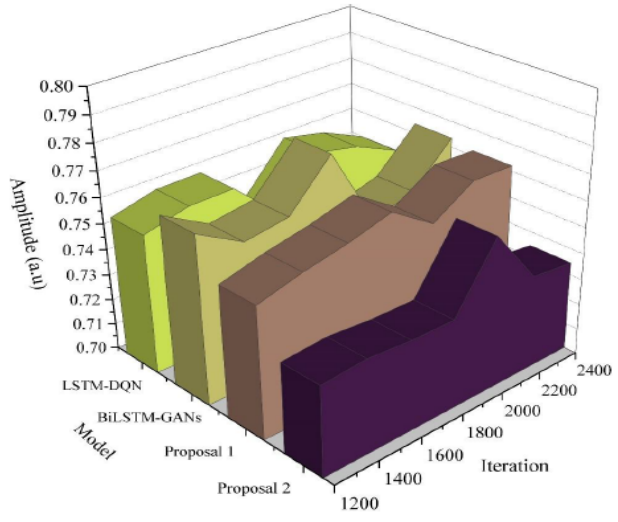
Backbone sound framework module test results are shown in Figure 4, the model function of this paper is to realize the guzheng playing style is to learn, and then the guzheng sheet music to rhyme, based on the current 60 sheet music selected 20 of them slightly similar style of the track for testing, the overall data set is divided into five for cross-validation testing. From the overall results of the test, the accuracy rate reaches 67.6% with a standard deviation of 2.34%, which is relatively satisfactory. Specifically, the number of spectral characters is 1 to 5, the average accuracy rate is higher, the standard deviation is smaller, and the model effect basically reaches more than 77% compliance rate, with an average accuracy rate of 80.28%. The number of spectral characters is 6 to 8, the average accuracy rate is lower, the standard deviation is larger, and the model effect is less satisfactory. The reasons are analyzed as follows, with spectral characters from 6 to 8, the sample set is smaller, which is unfavorable to both learning and testing, and when



(a)Soprano



(b)Alto



(c)Tenor

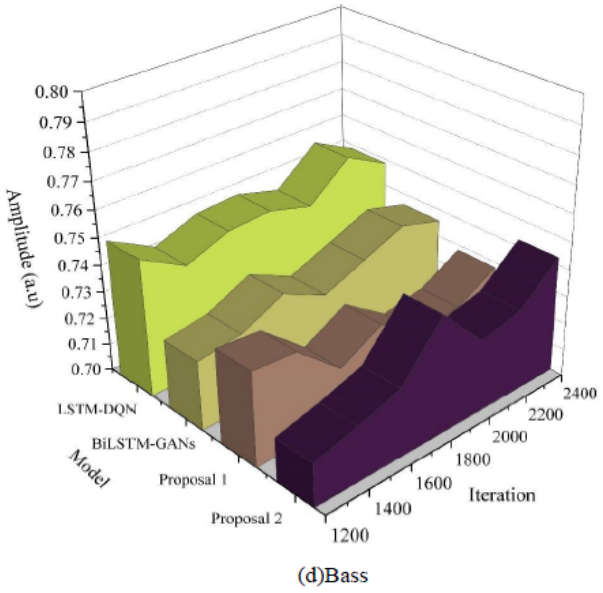


FIGURE 2. Comparison of model accuracy

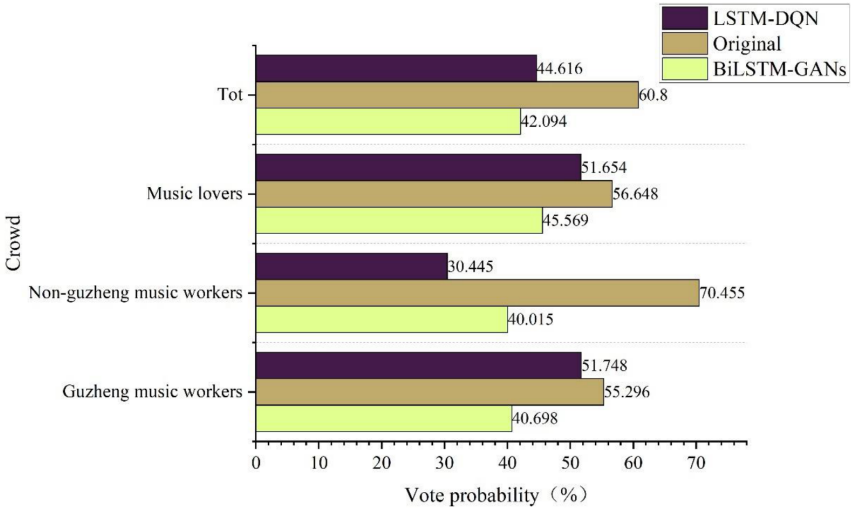


FIGURE 3. Results of manual evaluation

the spectral characters are from 6 to 8, there are more kinds of backbone rhythmic patterns, thus corresponding to a smaller training set, and thus the effect is less satisfactory. When the DQN model was added, a cross-validation test was performed using 16 pieces of music as the training set and 4 pieces of music as the test set, with an accuracy of 67.6% and a standard deviation of 2.34%, which is ideal. If there is a nonconformity in the rhyming score, the noise information is substituted into the rhyming score of the next bar, and this can lead to an increase in the nonconformity. Figure 5 shows the results of the amplification test, with an average accuracy of 64.38% and a standard deviation of 6.788%, which is relatively satisfactory. Specifically, the accuracy of the 16th note is the highest at 100%, which is supposed to be caused by the fact that there is only one case in which the actual note corresponding to the backbone of the 16th note is the 16th note. Secondly, the accuracy rate of eighth and quarter notes both reached more than 80%, 82.65% and 84.55% respectively, and the standard deviation was controlled under 5%, which is more ideal, which is related to the shorter note duration, less decomposition pattern, and more number of training sets, which have more influence on the whole rhyming score, and they account for the majority of the music ratio. Then there are the quarter and eighth appoggiaturas, which are also more effective, with an accuracy of more than 75% and a standard deviation controlled within 6%.

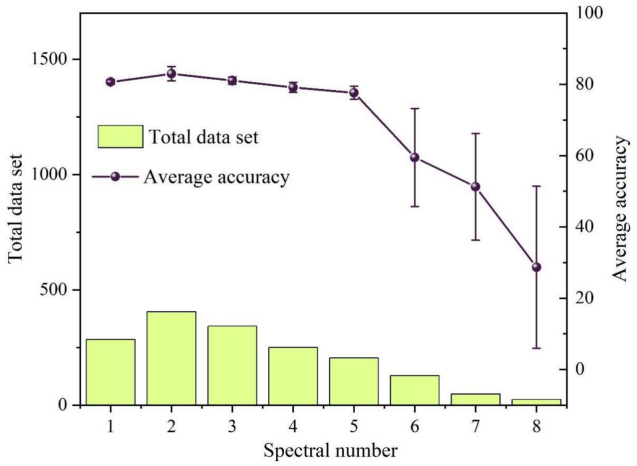


FIGURE 4. Key sound frame module test results

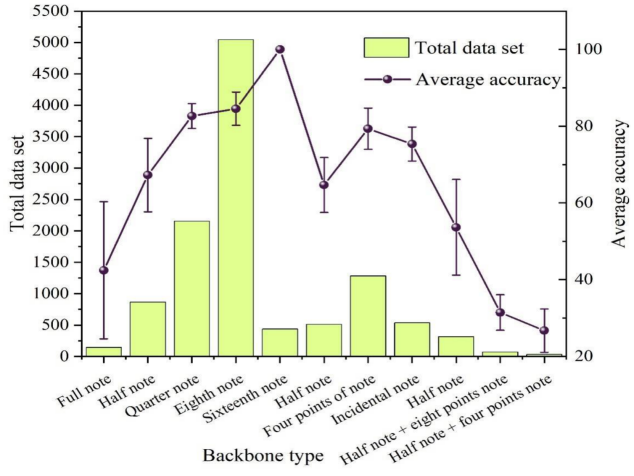


FIGURE 5. Amplification test results

V. CONCLUSION

In this paper, a similarity matrix extracts the main melody of guzheng music, identifying recurring segments via Python to define the chorus structure in a manner. These refined musical sequences are subsequently stored in the training database to facilitate the generation of harmonized compositions that mirror the rhythmic density of woven fabrics. The notes of guzheng music are generated by LSTM network, combined with the music theory and technique reward and punishment quantization mechanism of guzheng music, and the deep reinforcement learning DQN algorithm is used to generate the features of guzheng music, which facilitates the creation of guzheng music. According to the interactive functions of VR technology in auditory channel, physical touch, playing gesture and line-of-sight tracking, a virtual scene of guzheng performance is built. Evaluating the practice of zheng music creation and the aesthetics of composition, in the backbone sound frame test, the overall learning accuracy of guzheng playing style in this paper reaches 67.6%, with a standard deviation of 2.34%, which is more satisfactory, and is more prominent in the number of spectral characters 1-5, with an accuracy of more than 80%, which is able to provide good modeling technical support for the creation of the guzheng, so as to create guzheng compositions with a better artistic aesthetic. The model can provide good model technical support for guzheng creation, and thus create better artistic aesthetic guzheng arrangements. Scoring the guzheng compositions created by the model of this paper to assess the aesthetic value of them, the ratings of guzheng music professionals and non-guzheng professionals are relatively high, with the mean value of 3.04 and 3 points respectively, which is in line with the public aesthetics. Through the fusion of AI technology and VR technology, composers can rapidly transform ideas into specific zheng music works, greatly shortening the time cycle from conception to finished product and improving creative efficiency. This technological innovation makes the expressive power of zheng compositions significantly enhanced, whether it is a delicate emotional expression or a grand scene depiction, which can be more accurately and vividly presented by means of AI and VR.

AUTHOR CONTRIBUTIONS

Conceptualization – Liulingzi Xiao, Wenzhen Xiong and Haibo Zhang; methodology – Liulingzi Xiao, Wenzhen Xiong and Haibo Zhang; investigation – Liulingzi Xiao, Wenzhen Xiong and Haibo Zhang; writing-original draft preparation – Liulingzi Xiao, Wenzhen Xiong and Haibo Zhang. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This paper is one of the interim outcomes of the 2023 Changde College Key Project titled "Theoretical and Practical Research on Integrating Changde Opera Resources into the 'Introduction to Ethnic Music' Curriculum from the Perspective of Intangible Cultural Heritage Inheritance."

HUMAN RESEARCH SUBJECTS

This study was approved by the Ethics Committee of Changde College, and was performed in accordance with

the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] F. Kouwenhoven, "China: The Guqin zither," The other classical musics, pp. 105-125, 2015, doi: 10.1515/9781782045359-008.

[2] Z. Li, "Aesthetic Origin of the Metal and Stone Sound in the Chinese Seven-String Zither Music," Advances in Education, Humanities and Social Science Research, vol. 1, no. 1, pp. 53-53, 2022, doi: 10.56028/aehssr.1.1.53.

[3] I. Zinkiv and J. Ren, "CHINESE ZITHER SE IN ANCIENT AND TRADITIONAL MUSICAL AND INSTRUMENTAL CULTURES OF TURKISH, MONGOLIAN AND TUNGUSKA-MANCHU ETHNIC GROUPS," Yegah Müzikoloji Dergisi, vol. 7, no. 4, pp. 510-533, 2024, doi: 10.51576/ymd.1543588.

[4] X. Hu, "Technology in the Combination of Traditional and Modern Music Composition," In New Paradigm in Digital Classroom and Smart Learning: Proceedings of the International Conference on Digital Classroom and Smart Learning. New York, NY, USA: Springer Nature, vol. 54, pp. 294, 2024, doi: 10.1007/978-3-031-98607-9_29.

[5] J. Zheng, M. Cao, and C. Zhang, "AI-driven generation of guzheng music from classical Chinese poetry: toward a new paradigm of creative practice in Chinese traditional Music," Multimedia Systems, vol. 31, no. 6, pp. 437, 2025, doi: 10.1007/s00530-025-02023-w.

[6] S. R. Marshall, T. N. D. Tran, M. R. Tapas, and B. Q. Nguyen, "Integrating artificial intelligence and machine learning in hydrological modeling for sustainable resource management," International Journal of River Basin Management, pp. 1-17, 2025, doi: 10.1080/15715124.2025.2478280.

[7] A. Taweesan, T. Kanabkaew, N. Surinkul, and C. Polprasert, "Integrating clustering algorithms and machine learning to optimize regional snapshot municipal solid waste management for achieving sustainable development goals," Environmental Advances, vol. 19, Art. no. 100607, 2025, doi: 10.1016/j.envadv.2024.100607.

[8] C. Wang, W. Xiong, and G. Zhang, "Application of deep learning models with spectral data augmentation and denoising for predicting total phosphorus concentration in water pollution," Journal of the Taiwan Institute of Chemical Engineers, vol. 167, Art. no. 105852, 2025, doi: 10.1016/j.jtice.2024.105852.

[9] W. Hussain, M. F. Mushtaq, M. Shahroz, U. Akram, E. S. Ghith, M. Tlija, et al., "Ensemble genetic and CNN model-based image classification by enhancing hyperparameter tuning," Scientific Reports, vol. 15, no. 1, pp. 1003, 2025, doi: 10.1038/s41598-024-76178-3.

[10] X. Zhu, "RNN Language Processing Model-Driven Spoken Dialogue System Modeling Method," Computational Intelligence and Neuroscience, vol. (1), Art. no. 6993515, 2022, doi: 10.1155/2022/6993515.

[11] Y. Yang, "Performance skills and artistic conception of Chinese Zither music," Frontiers in Art Research, pp. 3(3), 2021, doi: 10.25236/FAR.2021.030304.

[12] Y. Kang, "Performance Techniques and Insights of the Guzheng Concerto "Chan Ge"," Journal of Education and Educational Research, vol. 9, no. 1, pp. 44-46, 2024, doi: 10.54097/m531pr78.

[13] Q. Su, "Contemporary Guzheng Inheritance Issues and Strategies," Arts, Culture and Language, pp. 1(10), 2024, doi: 10.61173/769vth96.

[14] Y. Kang and A. Sh, "An Initial Exploration of the Historical Development and Origin of the Guzheng," Highlights in Art and Design, vol. 6, no. 1, pp. 25-31, 2024, doi: 10.54097/gqgpm010.

[15] Y. Zhang, X. Lv, Q. Li, X. Wu, Y. Su, and H. Yang, "An automatic music generation method based on RSCLN_Transformer network," Multimedia systems, vol. 30, no. 1, pp. 4, 2024, doi: 10.1007/s00530-023-01245-0.