Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Research on Constructing English Speaking Teaching Scenarios Based on AI-Assisted Virtual Technology in Generative Adversarial Networks in the Intelligent+ Era

Jing Li[1,✉]

[1] *School of Foreign Languages, Wuhan College of Arts and Science, Wuhan, Hubei, 430345, China*

ABSTRACT

As artificial intelligence technology becomes more and more mature, it is both a challenge and an opportunity for English speaking teaching. Aiming at the poor generation of virtual English teaching resources due to the training problems of traditional generative adversarial network, dual generative adversarial network is used to optimize the above problems and select the virtual English teaching resources that meet the requirements with the help of Pielou. At this level, the HTC VIVE suite, high-performance computer system, Unity 3D development engine, and joystick control are integrated to jointly complete the work of English speaking teaching scene design. Combining the research data and evaluation indexes, the practical application efficacy of the scenario is analyzed. From the overall performance of different methods in the four datasets, this paper's method is superior to the other four methods, that is, this paper's method is able to generate high-quality virtual spoken English teaching resources. And the practical application efficacy in terms of test scores, learning effects, satisfaction, and English speaking teaching background is better than traditional multimedia, which is more conducive to promoting the development of English speaking teaching.

Keywords: generative adversarial networks, virtual technology, Unity 3D, spoken English teaching scenario

1. Introduction

In today's ever-developing globalization, the English language subject is becoming more and more popular in the society, the teaching requirements of the subject have become more stringent, and students' oral expression ability has been repeatedly mentioned. The new curriculum reform requires teachers to improve their teaching position, attach great importance to oral teaching, focus on cultivating students' awareness of cross-cultural communication, ensure that students look at Chinese and Western cultures rationally, analyze in depth the cultural roots behind linguistic knowledge, learn

✉ Corresponding author.
E-mail address: 15717159589@163.com (J. Li).
Received 05 March 2024; Revised 20 April 2024; Accepted 05 December 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-386

English knowledge with a good sense of cross-cultural communication as well as cultural self-confidence, develop strong oral expression skills, and update teaching materials in order to master the core essentials of English learning[1-3]. Teachers should rationally analyze the cognitive expectations and physical and mental development of students, explore new teaching paths based on the current reality of oral teaching, deeply explore the hidden oral teaching content in teaching materials, and step by step improve students' English thinking level and enhance their oral learning consciousness [4]. Language learning, especially the development of speaking ability, is more dependent on the external language environment, which is of great significance in the teaching of language subjects. The ultimate goal of English learning is to be able to use it freely in real life, therefore, a suitable educational scenario in the process of English oral learning, such as role-playing, situational teaching, interactive teaching, personalized learning experience, etc., can help students to master the oral skills of voice, intonation, speed, grammar, etc., and also enable students to build up self-confidence, enhance confidence, and promote self-development and growth [5-8]. Teachers should recognize that creating a good English speaking atmosphere is the key to unlocking the potential of students' oral expression in English, and the role of environmental influences can stimulate students' enthusiasm for oral expression in English, and at the same time can help students accumulate rich experience in oral expression through daily interaction and communication, so that they can break through the upper limit of the enhancement of their English speaking ability and obtain good development of language literacy [9-10]. Teachers should pay attention to the role of environmental education, create a strong English speaking atmosphere in the classroom, encourage students to take the initiative to express themselves and actively interact with each other, so as to improve the effect of English speaking teaching. However, this is not the case. In the current speaking classroom, the teacher's unilateral output, basic question and answer interaction, and the students' difficulty in speaking have led to a dull atmosphere in the classroom. And the teaching form is single, such as multimedia display, playing audio or video, etc., which lacks personalization and interaction, resulting in a low level of students' speaking ability [11-14].

With the development of Artificial Intelligence (AI) technology, English speaking teaching ushers in a new life. Generative Adversarial Network is a powerful machine learning technique consisting of a pair of competing neural networks, a generator network is responsible for generating virtual samples [15]. A discriminator network is responsible for distinguishing real samples from virtual samples and generating samples close to real data, which provides a new idea for the construction of English speaking teaching scenarios.

Generative adversarial networks have been put into application in English education. Literature [16] combined generative adversarial network and reinforcement learning to design a text generation model, which can assess students' translation ability and assist in improving their translation skill level in English translation teaching. Literature [17] designed a method for assessing and predicting the effectiveness of university English teaching through progressive recursive generative adversarial networks, and the accuracy of the evaluation was improved by 31.64% on average compared with the evaluation methods of augmented machine learning algorithm, Internet of Things, and association rule algorithm + machine learning. Literature [18] generates English cultural communication materials with the training generator of generative adversarial network, distinguishes between generated materials and raw materials with its judger, and generates materials with high click rate and participation rate in English cultural communication.

And the difference of English speaking teaching scene makes the students' learning efficiency differentiated. Currently constructed English teaching scenes are mostly situational simulations. Literature [19] simulates work scenes with pictures, videos, audio and other auxiliary teaching tools to teach nautical English listening and speaking, which breaks the boring teaching atmosphere and inefficient teaching efficiency. Literature [20] proposes a virtual teaching scenario of e-schoolbag, which is seamlessly connected with the real scenario, creates a good learning environment for

students, and significantly improves their English communication skills. Literature [21] uses virtual reality technology to simulate things, and with the help of VR glasses equipment and simulation applications, it is directly applied to the conversation learning in English teaching to exercise students' oral practice ability. Literature [22] set up the English speaking training mode of scenario courses, and the teaching effect of scenario courses is remarkable, and the expression of students' speaking skills is significantly improved. In addition, there are interactive teaching scenarios. Literature [23] assisted the teaching of spoken English with artificial intelligence chat speaking interaction, which not only improved the students' willingness to learn, but also the improvement of speaking skills is even more obvious.

This project proposes a method for generating English teaching resources based on Bi-generative Adversarial Network (BiGAN-VSG), which preprocesses the data and unifies the data scale to prevent the model from being unstable or difficult to converge during the training process due to excessive fluctuations in the data. The generated English teaching resources are selected using the Pielou index, so that the generated virtual English teaching resources can supplement the original English teaching resources, thus ensuring the high quality of the obtained virtual English teaching resources. On the basis of the virtual English teaching resources generated by the aforementioned generative adversarial network, the English speaking teaching scenario is constructed, and with the help of appropriate technical tools, the construction of the English speaking teaching scenario is completed, and the effectiveness of the scenario's application in the classroom is evaluated and analyzed.

2. BiGAN-VSG-based virtual teaching resource generation methodology

2.1. Generative Adversarial Networks (GAN)

2.1.1. Network structure. Generative Adversarial Network (GAN) is a powerful generative model, the basic principle of GAN is inspired by the two-player zero-sum game game, i.e., the total gain of two players is zero, but for their own benefit each player finds a way to gain what the other player has lost by causing the other player to suffer a loss [24-26].A GAN usually consists of a generator and a discriminator. The generator tries to generate new results with the same features as possible as the real data samples by learning the features of the real data samples. The role of the discriminator is to discriminate the generator and the formers from each other by means of a Nash equilibrium in order to distinguish the real samples from the formed samples as accurately as possible. To achieve this goal, it will be assumed that there are two types of participants: a generator and a discriminator. The generator determines the final result by extracting features from the actual data analysis, while the discriminator is responsible for determining whether the user input data information is derived from the actual data analysis. In order to achieve the ultimate victory, both parties need to continuously improve their models to enhance the generative capability and recognition. The optimization process lies in achieving the best results by regulating the participants' relationship with each other.

The computational and overall flow of GAN is illustrated through Fig. 1, where D is used to denote the generator, G is used to denote the discriminator, whereas here actual information x and random variables z are used as inputs. $G(z)$ denotes the sample generated by G, which follows p_{Data} the real data distribution, and if the discriminator D detects that it comes from the actual information x, it categorizes it as true and marks it as 1; conversely, if the input comes from $G(z)$, D it categorizes it as false and marks it as 0. The goal of D is to ensure the accurate categorization of the data source, whereas the goal of G is to make the performance of the generated number $\mathrm{G(z)}$ on D comparable to the performance of the actual information x performs comparable to the performance of the generator and the discriminator by means of adversarial optimization. When D the judgment ability reaches a certain level, but the data source cannot be accurately identified, it can be assumed that the generator G has successfully captured the distribution characteristics of the real data.

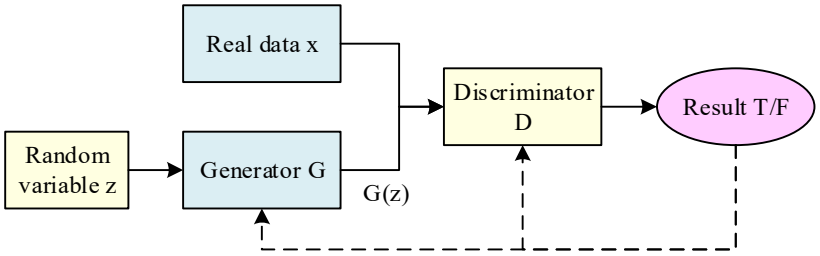


Fig. 1. Computation procedure and structure of GAN

2.1.2. Generators and Discriminators. For the training and testing of GAN networks in the first place the discriminator D of generator G is optimized to achieve the best performance. The classifier training method using Sigmoid parameters is similar, and the training of the discriminator also needs to consider the reduction of cross entropy. The loss function is expressed as follows:

$$Obj^{D}\left(\theta_D,\theta_G\right)=-\frac{1}{2}E_{x-p_{data}(x)}[\log D(x)] -\frac{1}{2}E_{z-p_z(z)}[\log(1-D(g(z)))] \tag{1}$$

where x is sampled from the true data distribution $p_{data}(x)$, z is sampled from the prior distribution $p_z(z)$, such as the uniform or Gaussian distribution, and $E(\cdot)$ denotes the expectation. Note that the training data consists of two parts: one from the true data distribution $p_{data}(x)$ and the other from the generated data distribution $p_g(x)$. This is slightly different from the traditional approach to binary classification. Given the generators, they need to minimize Eq. (1) to obtain the optimal solution. In continuous space, Eq. (1) can be reformulated as follows:

$$\begin{aligned} Obj^{D}\left(\theta_D,\theta_G\right)&=-\frac{1}{2}\int_x p_{data}(x)\log(D(x))dx \\ &-\frac{1}{2}\int_z p_z(z)\log(1-D(g(z)))dz \\ &=-\frac{1}{2}\int_x\left[p_{data}(x)\log(D(x))+p_g(x)\log(1-D(x))\right]dx \end{aligned} \tag{2}$$

expression for any $(m,n)\in R^2\setminus\{0,0\}$ and $y\in[0,1]$:

$$-m\log(y)-\log(1-y) \tag{3}$$

Thus, given generator G, Eq. (2) reaches its minimum under the following conditions:

$$D_G^*(x)=\frac{p_{data}(x)}{p_{data}(x)+p_g(x)} \tag{4}$$

The discriminator D of the GAN efficiently estimates the difference between two probability densities, which is the biggest difference from traditional Markov chain or lower bound methods that provide optimal solutions.

$D(x)$ represents the probability of extracting x from real statistics. When the input data is derived from real data x, the discriminator will try to approximate $D(x)$ to 1 as much as possible, whereas when the input data is derived from generated numbers $G(z)$, the discriminator will try to approximate $D(G(z))$ to 0, whereas the generator will try to approximate it to 1. Due to the existence of a zero-sum game between G and D and a loss function of G, $Obj^{G}\left(\theta^{G}\right)=-Obj^{D}\left(\theta^{D},\theta^{G}\right)$,

optimization of GANs can be generalized as a computation of a very small very large number, which effectively improves the generator's effectiveness and thus its optimization efficiency:

$$\min_G \max_D \left\{ f(D,G) = nE_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1-D(G(z)))] \right\} \tag{5}$$

In order to improve the accuracy of the GAN network model, the discriminator D needs to be trained to be able to accurately differentiate between the input data, the actual data analysis x, and the generated data analysis $G(z)$, in order to obtain the best results. In addition, the generator G needs to be trained to minimize the value of $\log(1-96)$. More advanced training methods can be used to improve efficiency and accuracy by adjusting the values of G and D to maximize the discriminative accuracy of D, and secondly, by adjusting the values of G and D to minimize the discriminative accuracy of D. This iterative process can achieve the global optimal solution at $p_{data} = p_g$. During the training process, the parameters of D are continuously adjusted and the parameters of G are adjusted once at a time to achieve optimal performance.

2.2. Fundamentals of BiGAN-VSG-based

In the face of the limitations of English teaching conditions and resources, the established English scenarios often face poor results due to the small sample problem and the traditional generative network training problem. To address this problem, an English teaching resource generation method based on bi-generative adversarial network (BiGAN-VSG) is proposed. Before introducing the BiGAN-VSG model, the preprocessing of data is firstly explained. The data scale is standardized to prevent the model from being unstable or difficult to converge during the training process due to excessive fluctuations in the data. Next, the introduction of Bi GAN-VSG contains two aspects, one is to train the BiGAN model by the original samples after preprocessing.BiGAN is able to generate the inputs and outputs of the English teaching resources directly and consider both the mapping relationship between the inputs and the outputs of the generated samples and the possible overfitting problem of training CGAN with small samples. When the training network reaches convergence, the Bi GAN model is utilized to generate the inputs and outputs of the English teaching resources to be selected. Second, the generated ELT resources are selected using the Pielou index to ensure that the virtual ELT resources can complement the original ELT resources. The constraints of the two steps ensure that the obtained virtual ELT resources are of high quality. Finally, the effectiveness of the Bi GAN-VSG method is evaluated by comparing the prediction accuracy of the Light-GBM regression model trained on the original English teaching resources and the modified English teaching resources.

2.2.1. Data pre-processing. Normalization is one of the important steps in the preprocessing stage of English speaking teaching resources data, which helps to eliminate the differences in scale between different features, making the training of the model more stable and accurate. In the process of generating virtual spoken English teaching resources, the data collected by different devices may have different scales, such as pressure, temperature, flow rate, etc., which will lead to large differences between the data and affect the training effect of the model. Therefore, it is necessary to normalize the samples. In this study the samples are mapped onto the $[0,1]$ interval by minimum one maximum normalization. The specific formula is as follows:

$$X = \begin{bmatrix} X_{II} & \cdots & X_{li} \\ \vdots & \ddots & \vdots \\ X_{jl} & \cdots & X_{ji} \end{bmatrix} \tag{6}$$

$$X'_{lk} = \frac{X_{lk} - X_{k_{\min}}}{X_{k\max} - X_{k\min}} \tag{7}$$

$$X' = \begin{bmatrix} X'_{ll} & \cdots & X'_{li} \\ \vdots & \ddots & \vdots \\ X'_{jl} & \cdots & X'_{ji} \end{bmatrix} \tag{8}$$

where l represents rows, $l \in \{1,2,\ldots j\}$, k represent columns, $k \in \{1,2,\cdots i\}$, X_{lk} is a data to be normalized, $X_{k_{\min}}$ is the minimum value of k columns, $X_{k_{\max}}$ is the maximum value of k columns, X is the original sample matrix, and X is the normalized sample matrix. The inverse normalization is to make the normalized data be converted back to the original range of values, thus retaining the original information of the data and making the data applicable to the actual needs of virtual spoken English teaching resources generation, with the following formula:

$$X_{lk} = X'_{lk} * \left(X_{k_{\max}} - X_{k_{\min}} \right) + X_{k_{\min}} \tag{9}$$

2.2.2. Loss function. For the VSG of regression tasks, traditional methods usually focus only on obtaining the input data of virtual spoken English teaching resources, and then obtaining the output of the virtual samples through the small-sample trained regression model. So much so that this approach ignores the importance of the output of virtual spoken English teaching resources and the potential mapping relationship between sample input and output, resulting in the accuracy of the soft measurement model failing to meet the desired requirements. To solve this problem, BiGAN model is proposed whose structure is shown in Figure 2, which consists of two discriminators and two generators. Among them, generator Gx negatively generates the input data (x'), generator Gy negatively generates the output data (y'), discriminator Dx is responsible for discriminating between the real and generated input data, and discriminator Dy negatively discriminates between the real and generated output data.The details of the BiGAN model are described as follows:

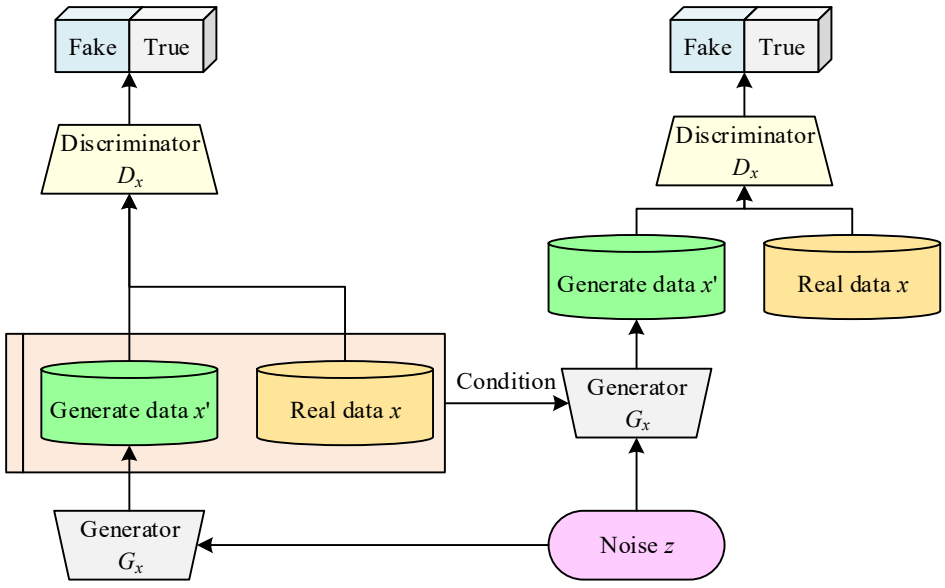


Fig. 2. The structure of BiGAN

The L2 loss is added to the loss function of the generator to constrain the generated virtual English speaking teaching resources to satisfy two requirements: firstly, the distribution similarity with the

real data, and secondly, the matching with the original mapping relationship between the input and the output. And the CGAN with L2 loss is named as RC-CGAN.By this way, RC-CGAN can better meet the requirements of the distribution similarity between generated samples and real data and the mapping relationship between inputs and outputs, which improves the performance and prediction accuracy of the model.The loss function of RC-CGAN is represented as follows:

$$\begin{aligned} \min_{L_{G_y}} &= E_{\substack{z\sim P_z(z)\\ x\sim P_{data}}} \lg lop\left(1-D_y\left(G_y(z\mid x)\mid x\right)\right) \\ &\quad +MSE_{x,y\sim P_{data}}\left(G_y(z\mid x),y\right) \end{aligned} \tag{10}$$

$$\max_{L_{D_y}} = E_{x,y\sim P_{data}} \lg D_y(y\mid x) + E_{\substack{z\sim p_z(z)\\ x\sim p_{data}}} \lg\left(1-D_y\left(G_y(z\mid x)\mid x\right)\right) \tag{11}$$

In addition to this, for deep learning prediction models, when the amount of training data is limited, the model is more likely to suffer from the overfitting problem, which affects its generalization ability and prediction performance. In order to solve this problem, a direct and effective method is to carry out sample expansion, i.e., generating additional virtual English speaking teaching resources data to enrich the training set, so as to improve the robustness of the model.WGAN-GP can generate virtual samples with the same distribution as the original samples and use them as inputs to the CGAN, but the generated data do not have labeling information, for which semi-supervised methods are used for training. The specific training process is:

First fix the generator and optimize the discriminator. When the condition is x and the input is y, the discriminator is able to determine it as true. At the same time, when the condition is x and the input is $G_y(z\mid x)$, the discriminator is expected to be able to judge it as false.

Then, fix the discriminator, optimize the generator in the hope that the discriminator will judge it true when the condition is x and the input is $G_y(z\mid x)$, and then adjust the generator.

Next, use the generator to predict the output $G_y\left(z\mid x'\right)$ of x, and optimize the generator when input $\left(G_y(z\left|x'\right)\right|x)$ to the discriminator, hoping that the discriminator will still judge it true.

The above steps are repeated until the loss converges. In this way, the data distribution learned by the generator of CGAN is expanded, which enhances the robustness of learning and makes the generated data closer to the real data. Thus the loss function of BiGAN is as follows:

$$\max_{L_{D_x}} = E_{x-p_{data}(x)}D(x) - E_{z-p_z(z)}D(G(z)) + \lambda E_{\hat{x}-p_{\hat{x}}}\left[\left(\left\|\nabla_{\hat{x}}D(\hat{x})\right\|_2 -1\right)^2\right] \tag{12}$$

$$\min^{L_{G_x}} = E_{z-p_z(z)}[D(G(z))] \tag{13}$$

$$\max_{L_{D_y}} = E_{x,y\sim P_{data}} \lg D_y(y\mid x) + E_{\substack{x\sim P_z(z)\\ x\sim data}} \lg\left(1-D_y\left(G_y(z\mid x)\mid x\right)\right) \tag{14}$$

$$\min^{L_{Gy}} = L_{G_y\,\mathrm{sup}ervised} + L_{G_y un\,\mathrm{sup}ervised} \tag{15}$$

$$\begin{aligned} L_{G_y\,\mathrm{sup}ervised} &= 0.5^* E^{x\sim P_{data}}_{z\sim P_z(z)} \lg\left(1-D_y\left(G_y(z\mid x)\mid x\right)\right) \\ &\quad +MSE_{x,y\sim P_{data}}\left(G_y(z\mid x),y\right) \end{aligned} \tag{16}$$

$$L_{G_y un\,\mathrm{sup}ervised} = 0.5^* E_{\substack{z\sim P_z(z)\\ i\sim \hat{P}_{gen}}} \lg\left(1-D_y\left(G_y(z\left|x'\right)\right|x)\right) \tag{17}$$

where $z\sim p_z(z)$ means that z obeys a standard normal distribution, $x'\sim P_{gen}$ means that x' is a sample generated by WGAN-GP, and $x,y\sim P_{data}$ means that x,y obeys the real data distribution.

When the network reaches the Nash equilibrium, the data (X', Y') generated by the model should conform to the same distribution as the real data (X, Y), and the mapping functions of the conditional inputs x to $G_y(z \mid x)$ and the mapping functions of the generated conditional inputs $\hat{x}$ to $G_y(z \mid \hat{x})$ should be approximately the same as the mapping functions of x to y. Ultimately, a certain number of generative samples (X', Y') to be selected can be generated by the BiGAN model.

2.2.3. Selection of virtual ELT resources based on the Pielou Index. In order to constrain the virtual English teaching resources data to fall into the interval of lower probability density, Pielou index is used as the basis for sample selection as follows:

Step1: Calculate the uniformity index (E_j) for each dimension:

1) Divide each dimension into n equal parts, count the number of English teaching resources falling into each interval, and calculate the frequency of each interval $p_i(p_i 0)$;

2) Calculate the uniformity index for each dimension using the Pielou index formula E_j:

$$E_j = \sum_{i=1}^{n} P_i * \ln P_i / \ln n \tag{18}$$

Step2: Calculate the average uniformity index $(\overline{E})$:

Assuming a total of b dimensions, calculate the average of the uniformity indices of all dimensions $\overline{E}$:

$$\overline{E} = \frac{l}{b} \sum_{j=l}^{b} E_j \tag{19}$$

Here the average uniformity index indicates the degree of uniformity of the overall data distribution, the larger indicates the more uniform data distribution.

Step3: Filter out the virtual samples that satisfy the conditions:

1) Use the trained BiGAN to generate samples, and generate a certain number (b) of virtual samples in each batch.

2) Calculate the average uniformity index $\overline{E}$ of the original samples, add the generated virtual samples to the original samples, and recalculate the uniformity $\overline{E_{+1}}$. If $\overline{E_{+1}} > \overline{E}$, it means that the uniformity of the data is improved after adding the virtual samples, and these virtual samples can be saved; otherwise, discard them.

3) Add the appropriate virtual samples to the original samples, continue to generate a new batch of samples and compare them with $\overline{E}_1$ without adding new virtual samples, still keeping the method of preserving the data if $\overline{E_{i+1}} > \overline{E}_i$.

4) Repeat the process until a sufficient number of target virtual samples are selected.

2.2.4. Evaluation criteria. The purpose of BiGAN-VSG is to generate virtual samples mapped similar to the original samples X, Y by BiGAN and inject them into the original samples, so as to improve the accuracy of the prediction of the soft measurement model. Next, is the process specific to BiGAN-VSE:

1) Normalize the inputs X and outputs Y of the original samples to obtain the normalized samples (X', Y').

2) Divide (X', Y') into training set (X'_{train}, Y'_{train}) and test set (X'_{test}, Y'_{test}), and use the training set to train the BiGAN model.

3) The BiGAN after training is completed generates both sample inputs and outputs, and virtual sample inputs and outputs $(X'^{vir}_{train}, Y'^{vir}_{train})$ are selected by the uniformity index described above.

4) Only the output data are back-normalized to obtain datasets $\left(X'_{train}, Y_{train}\right)$ and $\left(X'^{vir}_{truin}, Y^{vir}_{train}\right)$ and use them to construct the Light-GBM soft measurement model, and the virtual samples are used to verify the enhancement of the model accuracy using test set $\left(X'_{test}, Y_{test}\right)$.

5) The mean absolute error (MAE), root mean square error (RMSE), and root mean square error enhancement $\left(EIR_{RMSE}\right)$ are used as the error analysis of the prediction model, and their equations are as follows:

$$MAE = \frac{l}{m}\sum_{i=1}^{m}\left|y_i - \hat{y}_i\right| \tag{20}$$

$$RMSE = \sqrt{\frac{l}{m}\sum_{i=1}^{m}\left(y_i - \hat{y}\right)^2} \tag{21}$$

$$EIR_{RMSE} = \frac{RMSE - RMSE_{vir}{}^*}{RMSE} * 100\% \tag{22}$$

3. Construction of scenarios for teaching English as a foreign language

The development of information technology promotes the change of foreign language teaching concepts and teaching methods, and virtual reality technology and dual-generation adversarial network, as a new medium of information technology, bring new momentum to the development of foreign language teaching. How to utilize the advantages of virtual reality technology and very-generative adversarial network in language teaching, so that students can get a better learning experience, has become a problem that we need to pay attention to. On the basis of the virtual English teaching resources generated by the aforementioned generative adversarial network, based on the relevant theories of foreign language teaching and the scenario-based characteristics of VR, this subsection designs an oral English teaching scenario for oral English conversation training. This subsection elaborates the construction process of the spoken English teaching scenario.

3.1. Technical tools

On the basis of the virtual English teaching resources generated by the aforementioned Generative Adversarial Network, the construction of English speaking teaching scenarios requires the use of appropriate technological tools, and at the same time, the development of visualization programs also puts forward certain requirements on the ability to use computers. The hardware systems and development tools used in this study mainly include the following:

3.1.1. HTC VIVE Kit. The first thing you need to equip yourself with when using virtual reality technology is a VR device that can fulfill the functions of the system. There are already a large number of consumer-grade virtual reality technology tools available in the market with varying performance and functionality, such as the OculusRift, HTC VIVE, Windows Mixed Reality OEM headset, and so on. Among them, HTC VIVE has a high binocular resolution (2160×1200) and is less prone to vertigo. In addition to the two handheld controllers can capture the manipulator's movements, it also has a Valve patented control system positioning system - Lighthouse, relying on laser and photosensitive sensors to determine the position of moving objects, you can role position of the precise capture. In order to better restore the scenes and characters in the software design, and to realize the physical motion capture such as walking in the course design, the HTCVIVE kit was selected for development in this study.

3.1.2. High-performance computer systems. The computer running the virtual reality environment requires high graphics performance. In order to ensure the smooth and favorable operation of the

environment, you can refer to the HTC VIVE headset manufacturer's standards for the operation of the equipment to be selected, the processor for the Intel core i5-4680/ AMD FXTN 8350 or higher, the memory is 8GB RAM or more, the GPU for the NVIDIA® GeForce® GTX 4060, AMD Radeon RX500 or higher, video output is HDMI 1.4. HDMI 1.4, DisplayPort 1.2 or above for video output, HDMI 1.8, DisplayPort 1.6 or above for USB, and Windows 8 SP1 64bit or later for computer system.

3.1.3. Unity 3D Development Engine. Unity 3D developed by Unity Technologies is a development engine that supports virtual reality. It integrates a large set of tools for rendering, physics, sound, input, networking, and more, and is used to develop 3D video games, architectural visualizations, real-time 3D animations, and other types of interactive content.Unity3D integrates SDKs from the major VR headset vendors, and Steam VR plug-ins are available for free from the Unity Store, which contain encapsulated prefabricated parts for scripting the HTC VIVE camera and controller scripting.

3.1.4. Handle control. In the CameraRig prefab, the Controller component represents the handle, which is divided into a left-hand handle and a right-hand handle. The interactions in this system are relatively simple, so only a single handle is needed to complete the control. The handle interaction contains many options, and this system uses a touchpad to realize the interaction with the spoken English teaching scenario.

3.2. Scenography

Compared with traditional computer desktop learning, the most important feature of immersive VR is the high-fidelity English scene experience. This study combines the virtual English teaching resources generated by the Generative Adversarial Network to design a multisensory experience including to the visual, auditory, and tactile senses. Learners can interact with multi-senses in the virtual reality world to achieve a more immersive scene effect. The use of virtual English speaking scenarios in English teaching is not only in line with the law of language acquisition, but also adapts to the trend of the times; not only is it a brand-new experience for students, but also a pedagogical innovation for teachers. With the support of virtual simulation technology and generative confrontation network, the teaching content is no longer limited to the textbook, and the teachers can add pictures, videos and other materials related to the curriculum to the virtual simulation system, continuously improve the virtual spoken English teaching scene, and change the teaching activities from the traditional duck-filling to induction and interactive, so as to obtain better teaching results.

3.2.1. Visual design. Vision is the most important way for human beings to receive information, and the visual presentation effect is the decisive factor for virtual reality scenes. Based on the above generative English teaching content, this study designs and builds an English speaking teaching scene, which contains a virtual environment and virtual characters.

1) Virtual Scene. The environment involved in the English content generated by the above network is mainly the internal environment of the restaurant. The virtual environment includes the restaurant building, indoor lighting, dining tables, seats, and various small objects (menus, vases, red wine). The 3D files for the restaurant environment were purchased from the Unity Store, and the desktop objects (e.g., menus) were created in Rhinoceros and imported.

2) Virtual Character. The character contains the main character A, and the surrounding secondary characters (waiters, bartenders, dining guests, etc.) All character models are purchased from the Unity official mall, and all contain skin and bone nodes.

3.2.2. Auditory design. Sound effects can emphasize the atmosphere of the scene and enhance the realism of the scene. The audio-visual experience of this system complements each other, with the

virtual environment generating environmental sound effects and the virtual characters generating dialog sound effects to help users quickly integrate into the scene.

1) Dialog sound effects. In order to enhance the realism of the dialog, it helps learners to experience the tone of the dialog more directly, and at the same time helps to correct the learners' voice intonation.

In this study, foreign students were invited to dub the dialog content in its entirety, and only then were English speaking teaching resources generated and imported into the virtual reality environment. The triggering of the dialog audio file and the triggering of the character's actions maintain a correspondence, i.e., what is said is what is done. When the learner is controlling the dialog line, he/she can hear the main character A speak at the same time, and at the same time, he/she can also see the action of the main character A at that time, and this correlation enhances the realism of the VR environment.

2) Ambient sound effects. Environmental sound effects include: restaurant noises, restaurant music. Among them, the restaurant music adopts stereo sound to create a spatial sense of the environment for users. The stereo sound is set by the sound source component in Unity3D. The horizontal coordinate represents the distance and the vertical coordinate represents the volume. This system uses the simplest current volume change, i.e. the further away from the sound source the lower the volume.

3.2.3. Haptic design. According to the theory of physical action feedback (TPR), the interaction of body movements can assist language learning. Therefore, this system not only designs a set of movements for the avatar, but also guides the user to walk and sit down through the content. According to the TPR theory, the action feedback will help the subjects to learn and memorize the content of the conversation.

1) Walking. At the beginning of the dialog, A greets B to enter a barbecue chicken restaurant, and the learner needs to walk according to the instruction to enter the restaurant when A says “Come Here! In this process, the learner will get tactile feedback from the ground and the scene in front of him/her will change his/her perspective as he/she moves.

2) Sitting down. According to the English teaching content generated above, A guides B to take a seat at the dining table, and when the learner hears A say “Sit down please” in the VR environment, he needs to sit down against the chair (the chair is a pre-prepared scene prop) and do the action according to the instruction, which embodies the learning method of TPR body movement feedback. It is worth noting that this process needs to be done in a way that the learners can understand the content better. It is worth noting that this process needs to be supervised by assistant students to ensure the safety of the learners.

4. Exploratory analysis of English speaking teaching scenarios

4.1. Evaluation of the Generation Effectiveness of Virtual English Speaking Teaching Resources

4.1.1. Contrasting models. In this paper, four baseline methods are selected for comparison experiments with this paper's method, namely, GRU, Dr GAN, original IMR, and BiGAN without the introduction of VSG and WMSE loss functions.To ensure the fairness of the experimental results, the comparison methods and this paper's model are experimented with the same dataset and the same anomaly rate, and the experimental models are selected with the same hyper-parameters and the same training batch size, and the model generation effect is evaluated using a unified The experimental models are selected with the same hyperparameters and training batch sizes, and the model

generation effects are evaluated using uniform evaluation indexes to ensure the fairness of the experiments.

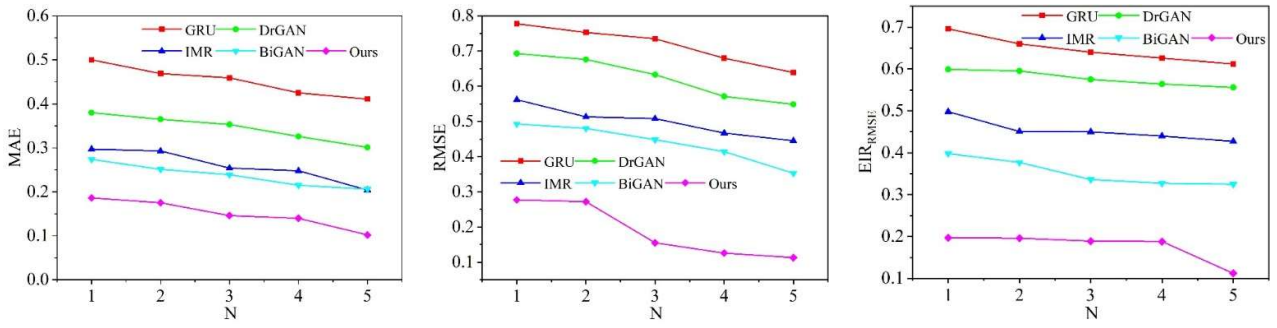
1) GRU: The temporal features of sequence datasets with different attributes are not the same, and the temporal dependencies of intervals in the English teaching resources dataset do not necessarily have rules to be extracted, the strong advantage of GRU network is to mine the temporal data, and the model that mines the temporal features according to a fixed time step does not necessarily obtain better generating results, so this paper takes GRU network as one of the comparison models.

2) DrGAN: This model is an English resource generation model based on improved generative adversarial network, the generator is mainly composed of up-sampling layer, convolutional layer and uses Leaky ReLU with tanh activation function, the discriminator is mainly composed of convolutional layer and uses Leaky ReLU with Sigmoid activation function, and the loss function adopts the basic RMSE.This model is the first time to introduce the generative adversarial network model into English resource generation.

3) IMR: This model et al. proposed a statistical-based teaching resource generation method for including English teaching resource data for information labeling.

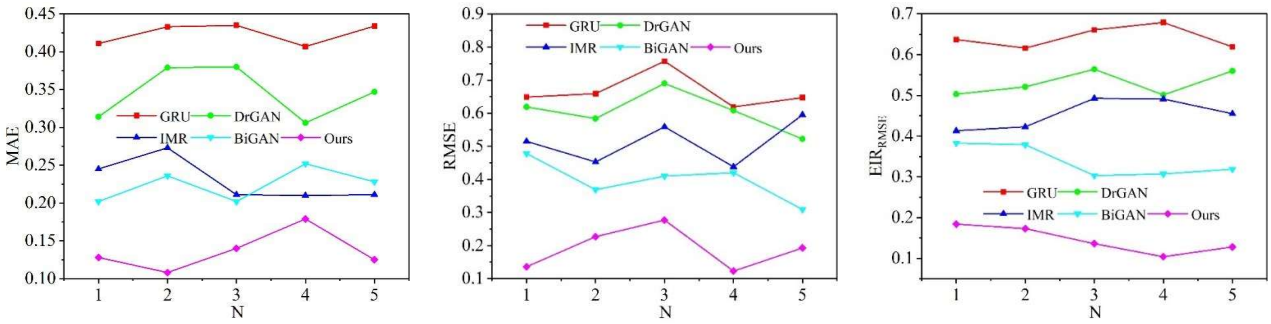
4) BiGAN: The model is based on the BiGAN model omitting the VSG, the network structure of the generator and discriminator are consistent with the network structure of the model in this paper, and retains the gradient penalty method, and the loss function adopts the reconstructed WMSE.

4.1.2. Results and analysis. In order to make the research data richer, this subsection selects one synthetic dataset (C), one open dataset (D), two English speaking datasets (E, F), all four datasets have been normalized, and experimentally compares the evaluation metrics (mentioned and defined above) of MAE, RMSE, and EIRRMSE of multiple generation methods including the algorithms of this paper, so as to verify that this paper's The generation accuracy and model stability of the methods are ahead of the comparative generation algorithm methods, and the evaluation of the generation effect of virtual English speaking teaching resources is shown in Figures 3~6, where (a)~(c) are MAE, RMSE, and EIRRMSE, respectively, the horizontal axis is the number of statistics, and the vertical axis is the index value. From the overall performance of different methods in the four datasets, compared with the other four generation methods, this paper's method has excellent application efficacy on virtual spoken English teaching resources, and its three index data are all lower than 0.35, which provides technical support for the design of spoken English teaching scenarios, so that the designed spoken English teaching scenarios are more in line with students' learning needs.



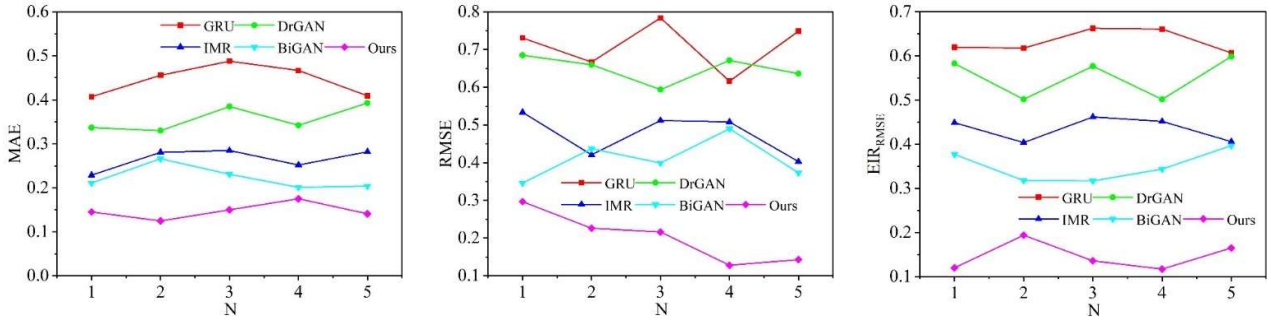
(a)MAE (b)RMSE (c)EIR$_{RMSE}$

Fig. 3. Synthesize data sets (C) to generate evaluation results



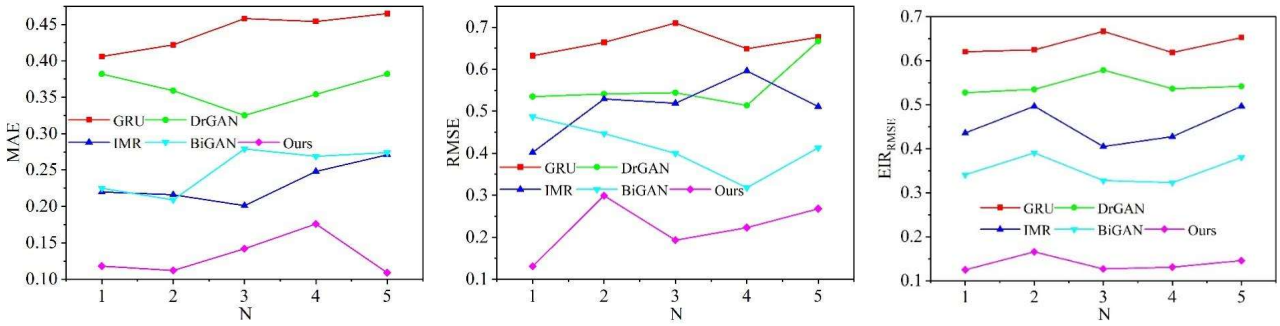
(a)MAE (b)RMSE (c)EIR_{RMSE}

Fig. 4. Synthesize data sets (D) to generate evaluation results



(a)MAE (b)RMSE (c)EIR_{RMSE}

Fig. 5. Synthesize data sets (E) to generate evaluation results



(a)MAE (b)RMSE (c)EIR_{RMSE}

Fig. 6. Synthesize data sets (F) to generate evaluation results

4.2. Analysis of Practical Effectiveness of English Speaking Teaching Scenarios

4.2.1. Objects of study. On the basis of the above VR+GAN-based English speaking teaching scenario, in order to verify its effectiveness and practical efficacy, the author conducted an application experiment in an elementary school in X city. The author chose a class of sixth grade in this elementary school (the class consisted of 62 students, 28 boys and 34 girls) to conduct the teaching experiment. In order to minimize the error of the experiment, the author, with the help of the English teacher of the class, divided the students of the class into two groups according to their English scores: the experimental group EG (31 students, 14 boys and 17 girls) and the control group CG (31 students, 14 boys and 17 girls). Due to the limitations of the school's normal teaching schedule and equipment, it

was decided in consultation with the school and the teachers that the experiment would be conducted with the two groups of students on Tuesday and Wednesday afternoons when the students were attending club classes.

4.2.2. Experimental variables:

1) Independent variables. Experimental group EG: Teachers use the VR+GAN-based English speaking teaching scenario to deliver lessons.

Control group CG: Teachers use the original, multimedia-assisted learning activity process for lectures.

2) Dependent Variables. The two groups in this experiment have different designs of speaking teaching activities, which may affect the learners' learning outcomes. The dependent variable in this experiment is the learning situation of the two groups of learners at the end of the study, which mainly includes the learning effect and the learning experience, and can be specifically divided into three dimensions: learning effect, learning interest and satisfaction with the learning activities.

4.2.3. Measuring tools:

1) Test Papers. At the end of the experiment, the “How can I get there?” unit test paper was distributed to all the subjects, and the differences between the two groups of students' mastery of the course were analyzed and summarized according to the completion of the recovered test papers. The design of the test paper was based on the integration of the requirements of the new standards for oral English teaching in elementary school and the teaching objectives of the learning activities, and examined the students' mastery of the content of the course, which mainly included listening questions, vocabulary questions, grammar questions, communicative questions and writing questions.

2) Questionnaire. In order to analyze the results of the data recovered from the questionnaire, the author rearranged the ideas and finally divided the dimensions of the questionnaire into the following three parts: learning effect (questions Q1~Q5 are, respectively, whether the fluency of oral expression is improved, whether the accuracy of pronunciation is improved, whether the level of confidence in communication is improved, whether the ability to apply the skills in practice is improved, and whether vocabulary use in the ability to expand the topic is improved), learning interest (questions Q6~Q10, respectively, are, respectively), and learning objectives (questions Q6~Q10, respectively, are, respectively). Q6~Q10: whether the attractiveness of the classroom content, the fun of classroom activities, the supporting effect, the atmosphere of peer interaction, and the teaching style have improved) and satisfaction with the teaching of spoken English (Q11~Q15: whether the teaching content is reasonable, the teaching method is reasonable, the teaching time is reasonable, the teaching equipment is reasonable, and the teaching activities are reasonable). (Q11~Q15 are, respectively, whether the teaching content is reasonable, whether the teaching method is reasonable, whether the teaching time arrangement is reasonable, whether the selection of teaching equipment is reasonable, whether the teaching activities are reasonable). The reasonableness of the questionnaire was examined by SPSS software, and it was found that the questionnaire had good reliability and validity performance, which ensured the credibility of the research results.

3) Interview outline. After the experiment, after-class interviews were conducted with the English teacher of the class and a few students of the experimental group (two male and two female students were randomly selected). The interview method in this experiment serves two main purposes: first, to understand the differences between the daily conversation skills of the students in the experimental group and the control group through the interviews with the teachers, and to supplement the data collected from the test papers and questionnaires. The second is to understand the feelings of the

interviewees during the experimental process and their opinions and suggestions on the design of the whole activity. The questionnaire and interview methods were combined to collect data and used as a basis to analyze the actual efficacy of this English speaking teaching scenario. In addition, the conclusions of the experiment are summarized to provide lessons for improving the software and learning activity design.

4.2.4. Experimental procedures. The main research method used in this study is the experimental method, in order to ensure the validity of the experimental results, it is necessary to control the irrelevant variables (such as: the teacher, the content of the lesson, etc.) to remain consistent. Due to the limitations of the conditions and equipment, the author finally decided to conduct the experimental operation on the experimental group and the control group in a sixth grade class of an elementary school at unequal times.

4.2.5. Analysis of experimental results:

1) Analysis of test paper results. A total of 62 valid test papers were recovered at the end of the experiment, including 31 in the EG group and 31 in the CG group. The results of the test papers recovered from the EG and CG groups were counted, and the results of the test papers were analyzed as shown in Figure 7. It can be seen that the average score of the students in the experimental group is 76.78 and the average score of the students in the control group is 70.14, which indicates that the average level of the students in the experimental group is slightly higher than that of the students in the control group. There are 5 students in the experimental group with scores of 90 or more, the same number as the control group, indicating that for students with excellent English learning performance, the use of GAN+VR-based English speaking teaching scenarios for lectures does not have a significant effect on their learning performance compared with the use of multimedia lectures. In the experimental group, there are 7 students in the 80-90 score range, while in the control group there are only 3 students in this score range, and the difference in the number of students in the 70-80 and 60-70 score ranges between the two groups is not significant. In the score band below 60, there are 3 and 4 students in the experimental and control groups, respectively. Overall, the use of GAN+VR-based English speaking teaching scenarios for lectures has a more significant effect on students with upper-middle and lower scores, and a smaller effect on learners in other score bands.

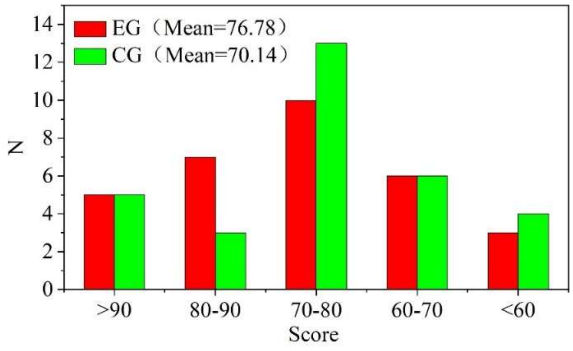


Fig. 7. Analysis of test volume results

2) Analysis of questionnaire results:

(1) Learning effect analysis. The comparative analysis of learning effect dimensions between the experimental group and the control group is shown in Figure 8, which shows that there is a significant difference between the experimental group and the control group in terms of the overall level of learning effect (P=0.044>0.05), and that the effect of using the spoken English teaching scenario for lectures is more obvious compared to multimedia-assisted teaching.

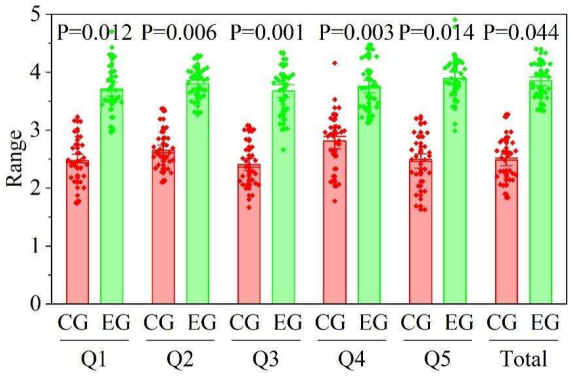


Fig. 8. Dimension comparison of learning effect between EG and CG

(2) Learning interest analysis. The comparative analysis of learning interest dimension between the experimental group and the control group is shown in Figure 9. There is a significant difference between the experimental group and the control group in the learning interest dimension ($P<0.05$), which shows that the learning interest level of students in the experimental group is significantly higher than that of students in the control group, i.e., it is easier to activate the students' learning interest in the English speaking teaching scenario.

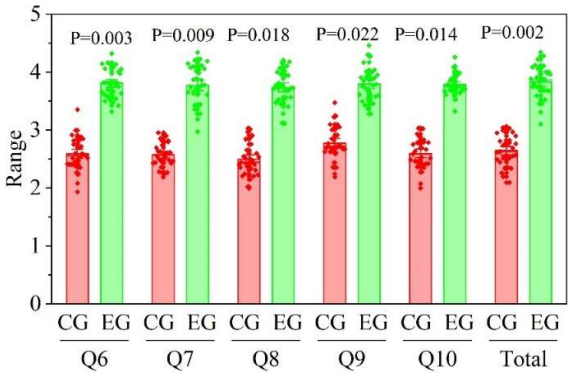


Fig. 9. Comparative analysis of learning interest between the EG and the CG

(3) Analysis of satisfaction. Figure 10 shows the results of the comparative analysis of the satisfaction dimensions of the two groups of students. Overall satisfaction $P=0.007<0.05$,proves that there is a significant difference between the experimental group and the control group students in the overall satisfaction dimension of the learning activities, and it can be seen that the overall level of the experimental group students' satisfaction with the learning activities is higher than that of the control group students, which not only verifies that this paper's method can effectively improve the students' satisfaction with the teaching of the speaking classroom, but also confirms the actual efficacy of the teaching of spoken English teaching scenarios.

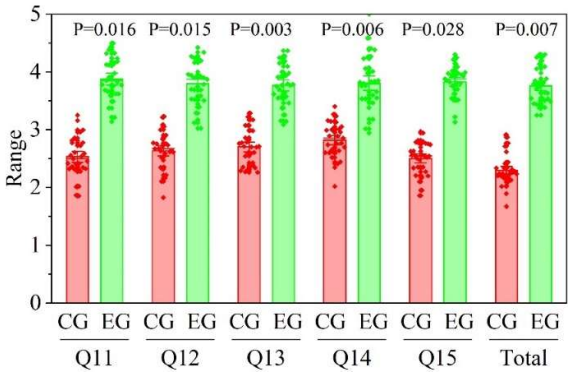


Fig. 10. Comparison of dimensions of satisfaction between the two groups of students

5. Conclusion

In view of the fact that traditional spoken English teaching is difficult to meet the current students' learning needs, BiGAN-VSG (Bi-Generative Adversarial Network) and Virtual Reality (VR) are used to jointly create a spoken English teaching scenario and explore the effectiveness of the practical application of this teaching scenario. Among the four datasets, the method of this paper has high superiority in virtual English teaching resources generation, which provides theoretical support for the design of spoken English teaching scenarios. In addition, comparing with multimedia teaching, students' interest in learning, learning effect, and classroom satisfaction are significantly improved by using the spoken English teaching scenario, which satisfies the $P<0.05$ significance condition and verifies the practical application efficacy of the spoken English teaching scenario.

Funding

This work was supported by the 2023 Hubei Scientific Planning Project: “Practical Research on ‘AI + VR’ assisted College Oral English Teaching in the ‘Intelligence +’ Era” (Project Number: 2023GB131).

It was also supported by its sub-project, the teaching (scientific) research project of Wuhan College of Arts and Sciences in 2024 (Project Number: 2024xzk01), titled “Research on the Path of AIGC-Driven Ideological and Political Empowerment in College English Curriculum Teaching”.

References

[1] Tseng, C. T. H. (2017). Teaching" Cross-Cultural Communication" through Content Based Instruction: Curriculum Design and Learning Outcome from EFL Learners' Perspectives. English Language Teaching, 10(4), 22-34.

[2] Robert, R., & Meenakshi, S. (2022). Rereading oral communication skills in English language acquisition: The unspoken spoken English. Theory and Practice in Language Studies, 12(11), 2429-2435.

[3] Wang, Y., Derakhshan, A., Ghiasvand, F., & Esfandyari, M. (2024). Exploring Chinese and Iranian EAP students’ oral communication apprehension in English: A cross-cultural mixed-methods study. System, 125, 103437.

[4] Tian, Y. (2023). Exploration of the Innovative Path of English Speaking Teaching Mode under the Background of Artificial Intelligence. Contemporary Education and Teaching Research, 4(12), 617-621.

[5] Li, R., Razali, F., Ismail, L., Muhamad, M. M., & Ma, X. (2024). Fostering autonomous learning in oral English through role play: An exploration in course setting. International Journal of Instruction, 17(3), 581-598.

[6] Zhao, J. (2021). An Analysis on the Application of Interactive Teaching Approach in College Oral English Class. Advances in Educational Technology and Psychology, 5(7), 145-153.

[7] Shi, R. (2021). The Role and Application of Situational Teaching in Teaching Chinese as a Foreign Language–A Case Study of Primary Oral English Class. Journal of Contemporary Educational Research, 5(10), 131-136.

[8] Wu, W. C. V., Chen Hsieh, J., & Yang, J. C. (2023). Personalizing Flipped Instruction to Enhance EFL Learners' Idiomatic Knowledge and Oral Proficiency. In Learning, Design, and Technology: An International Compendium of Theory, Research, Practice, and Policy (pp. 883-905). Cham: Springer International Publishing.

[9] Ege, F., & Dharma, Y. P. (2018). Creating a Good Atmosphere for Speaking Activity in the Classroom. Journal of English Educational Study (JEES), 1(1), 1-8.

[10] Eddie, A. A., & Aziz, A. A. (2020). Developing primary students' spoken interaction skill through communicative language teaching. International Journal of Academic Research in Business and Social Sciences, 10(2), 303-315.

[11] Liu, Y., & Qi, W. (2021). Application of Flipped Classroom in the Era of Big Data: What Factors Influence the Effect of Teacher-Student Interaction in Oral English Teaching. Wireless Communications and Mobile Computing, 2021(1), 4966974.

[12] He, B., Guo, S., Chen, Q., & Rivera, H. (2022). Effects of school environment, classroom instruction, and self-efficacy on Chinese students' motivation for oral English. English as a Foreign Language International Journal, 2(4), 5-26.

[13] Utami, V. (2019). Teacher's Difficulties in Teaching Oral Communication Skills in Indonesia: A Comparative Literature Review. Linguists: Journal of Linguistics and Language Teaching, 5(2), 16-28.

[14] Xie, Q. (2020). Teaching English oral communication for China's English minor undergraduates: barriers, challenges and options. The Asian Journal of Applied Linguistics, 7(1), 73-90.

[15] Jabbar, A., Li, X., & Omar, B. (2021). A survey on generative adversarial networks: Variants, applications, and training. ACM Computing Surveys (CSUR), 54(8), 1-49.

[16] Zhang, L. (2022). An assisted teaching method of college English translation using generative adversarial network. Mobile Information Systems, 2022(1), 5408309.

[17] Adversarial, R. G. (2024). Assessment and Prediction of University English Teaching Effect Using Progressive Recurrent Generative Adversarial Network. J. Electrical Systems, 20(3s), 2335-2346.

[18] Wen, Y., & Meng, J. (2024, December). Diversified Generation and Evaluation of English Cultural Communication Materials Based on Generative Adversarial Networks (GANs). In 2024 IEEE 4th International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA) (Vol. 4, pp. 636-640). IEEE.

[19] Wang, C. (2018, July). The Application of Scene Design in Maritime English Listening and Speaking Teaching. In 2018 4th International Conference on Economics, Social Science, Arts, Education and Management Engineering (ESSAEME 2018) (pp. 515-518). Atlantis Press.

[20] Hu, L. (2021). Improvement of English Communication Ability With Virtual Scenes Based on Electronic Schoolbag. International Journal of Emerging Technologies in Learning, 16(13).

[21] Luo, X. (2022). Practice of artificial intelligence and virtual reality technology in college English dialogue scene simulation. Wireless Communications and Mobile Computing, 2022(1), 4922675.

[22] El-Naggar, B. E. E. S., Al-Hadi, T. M., & Mohammad, W. H. M. (2019). A Scenario-Based Program for Developing English Oral Expression Skills of Secondary Schoolers. Journal of Research in Curriculum Instruction and Educational Technology, 5(2), 159-174.

[23] Fathi, J., Rahimi, M., & Derakhshan, A. (2024). Improving EFL learners' speaking skills and willingness to communicate via artificial intelligence-mediated interactions. System, 121, 103254.

[24] Ying Liu,Xiaohua Huang,Liwei Xiong,Ruyu Chang,Wenjing Wang & Long Chen. (2025). Stock price prediction with attentive temporal convolution-based generative adversarial network. Array100374-100374.

[25] Haoyu Ge,Longfei Yin,Xikang Cui,Lingyun Zhu,Lei Chen & Guohua Wu. (2025). Moving target ghost imaging based on wasserstein generative adversarial networks. Optics and Laser Technology112542-112542.

[26] Sumit Saha,Krishn Katyal & Surendra Nadh Somala. (2025). Deep learning-based imputation framework for bridge health monitoring using generative adversarial networks. Knowledge-Based Systems113088-113088.