From sound to embodiment: AI sound imitation technology driven by N/CM model enables immersive communication:Take the first person monologue narration video as an example

Xinyu Zhu , Haowei Guan , Cong Zhang*

School of Journalism and Communication, Beijing Institute of Graphic Communication, Beijing, 102600,China

***Corresponding author, E-mail:zhangcong@bigc.edu.cn**

Abstract

Recently, video commentaries featuring first-person monologues have gained popularity online. These videos use AI sound technology to reconstruct the story world through the subjective perspectives of characters in the drama, allowing viewers to 'embody' the characters and engage in immersive storytelling. This study employs multimodal analysis (Research 1) and controlled variable experiments (Research 2), based on the narrative-coordination model (N/CM, Narration/Coordination Model). It sets up a scenario experiment with two commentary perspectives (first-person and third-person) and two user technology acceptance levels (high-tech and low-tech users). By analyzing the dimensions of sound and visuals in first-person monologue commentaries, the study explores how AI sound technology endows characters with vivid 'voice life' (Research 1). It also reveals the unique advantages of first-person narration in narrative depth, emotional resonance, and audience interaction (Research 2), aiming to explore the feasibility of using AI sound technology to enhance the film and television industry and to create embodied immersive experiences.

Keywords: AI onomatopoeia; N/CM model; first-person monologue narration; immersion; embodied cognition

1 Introduction

In the global context of actively integrating AI technology with the cultural industry, since 2016, the State Council has issued the ‘13th Five-Year National Science and Technology Innovation Plan,’ which prioritizes artificial intelligence technology. In 2017, the ‘New Generation Artificial Intelligence Development Plan’ elevated AI technology to a national strategy, with the film and television industry becoming a key focus. In 2021, the National Film Bureau released the ‘14th Five-Year China Film Development Plan,’ which explicitly promotes the application of AI, machine learning, and other technologies across the entire film industry chain. These three policy iterations have not only deepened the integration of AI technology with the film and television industry but also, with the reduction in computing costs and the widespread use of open-source tools, broken down the elitist barriers of traditional film and television industries, allowing ordinary people to gain professional-level production capabilities. A prime example is the first-person narration videos for TV dramas and films, which enhance the audience’s immersive experience and emotional connection by cloning the voices of characters and generating personalized scripts, allowing viewers to follow the first-person perspective of the characters to understand the story and its historical context. According to data from platforms like TikTok and B Station, such videos can achieve tens of millions of views per post, and the number of followers on these accounts has seen a significant increase.

This study focuses on the integration of AI sound technology with first-person narration, aiming to address the following key questions: (1) How does AI sound technology endow characters in the drama with a ‘voice life’? (2) What unique advantages does first-person narration offer in terms of narrative depth, emotional resonance, and audience engagement? (3) Beyond the film and television industry, can AI sound technology serve as a viable approach for embodied immersive experiences? (4) How can we define and regulate the scope and methods of applying AI sound technology?

2 Literature review

2.1 Immersion

Immersive experience (Immersion) was initially seen as an objective attribute of technical systems. According to the ‘Framework for Immersive Virtual Environments’ (FIVE) proposed by scholars Slater and Wilbur (1997), immersive experience can be objectively measured through technical parameters, including the Inclusivity, Extensiveness, Surrounding, and Vividness of the display system. This technology-centric perspective emphasizes that when a system fully covers the user’s sensory channels (such as vision, hearing, and touch) and achieves an interactive loop, it can achieve the highest level of immersion. However, with the advancement of cognitive science, scholars have begun to focus on the subjective psychological aspects of users. Scholars Witmer and Singer (1998) found a positive correlation between immersive experience and task performance and individual cognitive tendencies, indicating that immersive experience is the result of’ selective attention allocation ‘and’ environment-induced cognitive engagement, ‘essentially reflecting the user’s active inhibition of real-world perception. The embodied cognition theory proposed by scholars Schubert et al. (2001) further deepens this understanding, suggesting that immersive experience is composed of two core components: SpatialPresence and Involvement. Spatial presence refers to the user’s ability to construct a psychological model of the virtual space, while involvement reflects the user’s cognitive and emotional engagement with virtual events. In recent years, research on immersive communication has sought to integrate the dual perspectives of ‘technology and psychology.’ Scholars Cummings and Bailenson (2015) and Agrawal et al. (2019) have redefined immersion as ‘an individual’s temporary detachment from real perception due to deep cognitive engagement in a specific context, ‘using meta-analysis techniques to examine immersion differences from four dimensions: system, content, environment, and individual. Building on this, Cao Zhihui et al. (2024) introduced the cultural dimension for the first time. In their study of the’ Chang ‘an Twelve Hours’ district, they found that the use of localized Chinese symbols, such as Tang Dynasty architectural styles and street vendors ‘cries, can increase user immersion by 32%. The application of collective memory empathy helps construct users’ immersion to some extent.

2.2 Narrative/coordination model (N/CM model)

The Narrative Coordination Model (N/CM) is rooted in the interdisciplinary integration of narrative immersion mechanisms. Busselle and Bilandzic (2008) were among the first to explore the cognitive aspects of narrative immersion, proposing the ‘Narrative Understanding and Participation Model.’ This model revealed that immersion is generated through two parallel processing pathways: narrative and coordination, establishing the prototype framework for the N/CM model. However, it did not clearly define the neural mechanisms or the interaction between modality. Building on this, Cohen introduced a sensory dimension, proposing the ‘Harmonization and Synergy Model’ (CAM). He argued that narrative immersion is a dynamic balance between structural encoding and meaningful associations. Although this innovation bridged the gap from cognitive description to neural mechanisms, it still lacked an operational definition for perspective transformation. It was not until Denisova et al. (2015) approached the topic from a gaming perspective, comparing players’ preferences for game experiences, psychological structures, and performance forms, that the importance of perspective transformation in immersive experiences was truly elucidated. Thus, the N/CM model has provided a clear theoretical framework for understanding and researching immersive communication.

2.3 AI sound imitation technology and immersion

As the concept of immersive communication integrates into the core of media culture production, promoting the integration of media content into a third space that blends the virtual and the real, and optimizing the layout of the immersive communication industry has gained significant attention from the state. In 2020, the Ministry of Culture and Tourism released the ‘Opinions on Promoting the High-Quality Development of the Digital Cultural Industry,’ which stated, ‘Guide and support the application of technologies such as virtual reality, augmented reality, 5G + 4K/8K ultra-high-definition, and drones in the cultural sector, promote the transformation of existing cultural content into immersive content, and enrich virtual experience content. This indicates that immersive communication is not only a potential solution to meet people’s cultural and entertainment needs but also a key measure for achieving a modernization where material and spiritual civilization are in harmony. How to leverage AI technology to enhance immersive communication is an urgent issue that needs to be addressed. Based on the research subject of this article, the types of AI technology currently used in the market to enhance the immersion of films and television are as follows (see Table 1):

Table 1 Types of AI technologies used to enhance film and television immersion

Technology type	Core functions	technological superiority	The extent to which immersion is enhanced (Improved over traditional technology)
AI sound imitation technology	Voice cloning/ spatial audio synthesis	The auditory cortex responded at a speed of 150ms, and the amygdala was activated 2.3 times stronger	40-45% (spatial audio H=145.9)
AI image production	Image transformation/ scene generation	The production cost is reduced by 80%, and the visual impact MOS is 4.5	25-30% (resolution H=34.7)

AI real-time effects	Physical simulation/ dynamic interaction	The speed of iteration of the effect is increased by 10 times, and the cost of modification of the effect is reduced by 95%	35-38% (real-time feedback F=29.5)
Audio AR technology	3d sound field/ environmental sound synthesis	The cost of sound scene construction is reduced by 73%, and the fluctuation of sound pressure level ±6dB strengthens the decision-making pressure	30-33% (compared to traditional audio)
XR rendering technology	Multimodal fusion/ virtual reality interaction	The boundary ambiguity of MR environment virtual and real is 89%, and the delay of user action mapping is 12ms	28-32% (FOV expansion H=182.2)

The chart shows that the AI technologies currently used to enhance the immersive experience in film and television can be categorized into five main types. Among these, AI sound simulation technology has the highest impact on enhancing immersion, with a 40-45% increase. Additionally, sound-based technologies offer superior immersive effects compared to image and visual technologies. According to Greenwood (2003), the 'dual processing theory' suggests that sound can directly modulate attention allocation through non-conscious channels, processing information three times faster than vision. This allows soundscapes to more effectively guide users 'cognitive focus. Sanchez-Vives et al. (2005) conducted neuroscience experiments that confirmed that the spatiotemporal consistency of auditory stimuli triggers place illusions, significantly enhancing the audience's perception of spatial positioning. Building on the advantages of auditory elements in physiological and spatial dimensions for enhancing immersion, Slater et al. (2010) conducted a 'virtual environment simulation experiment.' They found that when the immersion level of the visual system decreases, users can still maintain 82% of their sense of presence through spatial audio compensation mechanisms. Conversely, if 3D sound effects are turned off, even the 4K panoramic visual immersion score drops by 63%. Auditory elements play a crucial role in constructing immersion in virtual environments. The experiments conducted by scholar Bansos et al. (2004) also demonstrated that AI-generated dynamic environmental sounds can enhance users 'spatial presence ratings, with the intensity of emotional arousal showing an inverted U-shaped relationship with the complexity of AI-generated sounds. The' dimensional sounds' created by AI significantly enhance immersion and spatial perception.

From the current AI sound simulation technology, AI sound simulation primarily uses deep learning models to deconstruct and reconstruct acoustic features. This process involves three stages: (1) extracting sound features using a CNN-RNN hybrid model, capturing static characteristics such as timbre and fundamental frequency, as well as temporal dynamic features; (2) employing a Transformer architecture for sequence modeling, using self-attention mechanisms to model long-term dependencies, thereby enhancing the naturalness of speech; (3) integrating adversarial training into the generative network (GAN), optimizing the realism of sounds through the dynamic interaction between the generator and discriminator. Today's AI sound simulation technology has achieved the capability of small-sample transfer learning, allowing for timbre cloning in just 3 to 10 seconds, with no restrictions on configuration, scenarios, or professionalism. This has enabled AI sound simulation technology to truly penetrate the market and cater to personalized needs. However, this progress has also brought about social ethical issues, such as AI sound simulation fraud.

Based on previous research, most studies on AI-empowered film and television production have focused on visual topics such as AI-generated images and short AI dramas, with fewer studies exploring the use of AI to enhance immersive experiences. Future research will focus on the N/CM model, examining how AI sound technology can enhance user immersion by constructing immersive virtual landscapes through sound. The following hypotheses and models are proposed (see Figure 1):

H1: The application of AI paronomasia monologue narration can predict the degree of user immersion positively.

H2a: AI paronomasia monologue narration can predict the degree of user immersion positively under the regulation of first-person narrative perspective.

H2b: AI paronomasia monologue narration can predict the degree of user immersion negatively under the adjustment of third-person narrative perspective.

H3a: The user's technical acceptance can positively predict the degree to which AI paronomasia monologue narration is applied.

H3b: Users' technology acceptance can positively predict their level of immersion.

H4: The degree of user immersion can positively predict the user's technology acceptance.

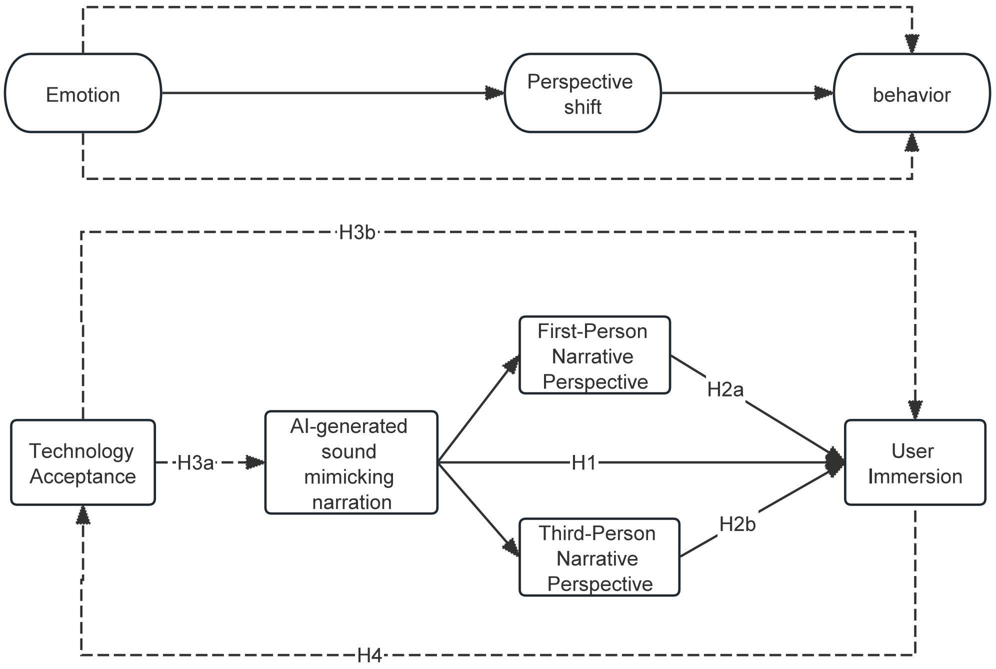

Emotion
Perspective shift
behavior
H3b
First-Person Narrative Perspective
H2a
Technology Acceptance
H3a
AI-generated sound mimicking narration
H1
User Immersion
Third-Person Narrative Perspective
H2b
H4

Figure 1 Hypothetical model of user immersion based on AI sound imitation technology under N/CM model

3 Research methods and objects

3.1 Multimodal analysis

Scholar Zhang Delu categorizes multimodal analysis into two dimensions: 'language' and 'non-language,' with the latter focusing on visual, bodily, and other forms of discourse expression. Given that this study focuses on AI paralinguistic technology, the analysis is limited to the 'language' dimension. A sample of videos featuring TikTokAI-generated first-person monologues with over 100,000 likes was selected for analysis, examining both the audio and text dimensions of the intertextual mechanisms. In terms of audio, Praat software was used to extract acoustic features such as fundamental frequency, formants, speech rate, and intonation from AI-generated speech. For the text, Nvivo qualitative coding tools were employed to analyze the emotional tendencies in generated dialogue texts, including psychological metaphors and the use of modal particles in character monologues. The study examines how AI paralinguistic technology constructs a embodied experience of the character's' voice life' from three dimensions: sound reproduction accuracy, emotional appropriateness, and contextual fit.

3.2 Control variable experimental method

This study designs an experimental design with two interpretation perspectives (first-person and third-person video commentary) and two user technology acceptance groups (high-tech acceptance group and low-tech acceptance group). Fifty highly active video commentary viewers, aged 18-35, are randomly assigned to form two groups: 25 in the high-tech acceptance group and 25 in the low-tech acceptance group. A two-way ANOVA model is constructed to test whether first-person narration has a compensatory effect on narrative advantages in the low-tech acceptance group. The experiment consists of three stages: pre-test (technology acceptance survey), intervention (watching randomly assigned video samples), and post-test (multi-dimensional scale + subjective interviews). The post-test scale is adapted from the Tsinghua University Virtual Reality Immersion Scale (2024), which evaluates users' immersion in VR systems through 17 dimensions, including sensory coordination and device usability, to optimize the system and enhance user experience. Given that the research targets the audience of AI-simulated sounds and the original questionnaire had overly complex measurement dimensions, this study revised the measurement objects and dimensions of the original questionnaire. The 'Virtual Reality (VR) system' was replaced with 'AI-simulated sound virtual space.' The immersion experience provided by AI-simulated sounds to users was measured from five dimensions: system level, content level, environment level, individual level, and cultural level. The measurement indicators included 30 items such as 'My emotions change with the AI-simulated sounds' and 'I gain new perspectives and insights through the AI-simulated first-person monologue.' A 5-point scale (1=not at all, 5=very much) was used, with higher scores indicating a greater sense of immersion. In this study, the Cronbach'sα coefficient of the scale was 0.900.

4 Research analysis and conclusion

4.1 Study 1: Multimodal analysis of AI paronomasia first-person monologue narration video

4.1.1 Descriptive statistics

To comprehensively gather target samples, this study entered various keywords such as 'first-person monologue narration,' 'AI monologue narration,' and 'open from the perspective of XXX' into the TikTok search bar, resulting in a total of 388 samples. After cleaning the sample videos based on their titles and tags, 368 valid samples were obtained, achieving an effective rate of 94.8%. Subsequently, AI-sound first-person narration videos with over 100,000 likes were selected for multi-modal analysis, covering multiple languages including Chinese and English, and genres such as film, television, history, and literature, forming an 8-multi-modal corpus. The basic situation of the crawled samples is as follows (see Table 2):

Table 2 Descriptive statistical analysis of the sampled samples

Video type	Average video length	More than 10,000 likes		10,000 likes		100,000 likes		Millions of likes		amount to	
		frequency	proportion	frequency	proportion	frequency	proportion	frequency	proportion	frequency	proportion
film and television	10min46s	253	81.9%	35	11.3%	20	6.5%	1	0.3%	309	84.0%
literature	10min22s	4	57.1%	2	28.6%	1	14.3%	0	-	7	1.9%
Comic and Animation	5min	7	77.8%	2	22.2%	0	-	0	-	9	2.4%
teach school	2min39s	13	92.9%	1	7.1%	0	-	0	-	14	3.8%
society	5min51s	4	80%	1	20%	0	-	0	-	5	1.4%
history	12min28s	4	80%	0	-	1	20%	0	-	5	1.4%

game	4min82s	6	100%	0	-	0	-	0	-	6	1.6%
figure	5min44s	11	84.6%	2	15.4%	0	-	0	-	13	3.8%
amount to	7min9s	302	82.1%	43	11.7%	22	5.9%	1	0.3%	368	100%

The chart shows that the first-person narration videos using AI sound simulation technology can be categorized into eight types, including film and television, literature, and animation. Among these, film and television videos, with 309 examples, account for 84.0% of the total, indicating a broader audience base and higher appeal for dramatic narratives. Despite the smaller sample sizes in animation, games, and social categories, these areas are also potential directions for future AI sound simulation technology applications. Additionally, it is noteworthy that the average duration of videos is inversely proportional to their like counts. Specifically, the average duration of first-person narration videos in film and television, literature, and history categories exceeds 10 minutes, yet they attract tens of thousands (22) and hundreds of thousands (1) likes within the target sample. This suggests that more in-depth video topics are better suited for AI sound simulation storytelling and enhance user immersion more effectively.

4.1.2 Multimodal discourse analysis

(1) Sound cloning dimension: measurement of basic acoustic indicators

This section selects the emotional climax segments from the Chinese AI parodied video “Revisiting ‘Painted Skin 2’ from Xiao Wei’s Perspective” with one million likes and the foreign AI parodied video “Telling the Story of American Psychopaths from Patrick Beterman’s First Person” with ten thousand likes as comparative cases. The software Praat is used to measure basic acoustic indicators such as pitch, sound quality, and rhythm. The aim is to explore the accuracy of AI parodied technology in reproducing the original characters’ voices and its potential for multilingual use in cross-cultural communication.

Table 3 Comparison of basic acoustic index measurement

Compare groups	Select the object	Pitch features				Sound quality characteristics			prosodic features
		Mean fundamental frequency	fundamental frequency min	fundamental frequency max	reduction ratio	Resonance peak frequency F1	Resonance peak frequency F2	Resonance peak frequency F3	Speech rate: words/s
Comparison 1: Original sound restoration comparison	Case 1	165.98	55.98	238.69	75.50%	722.16	1903.64	2975.87	5.2
	Case 2	218.83	144.78	351.16		760.91	1762.62	3014.63	4.57
Comparison 2: Cross-language AI sound effect comparison	Case 1	165.98	55.98	238.69	\	722.16	1903.64	2975.87	5.2
	Case 3	101.58	61.76	408.41	\	763.56	1772.23	2846.39	3.4

Note: Case 1: "Re-telling 'Painted Skin 2' from the Perspective of Xiao Wei" with an onomatopoeic version; Case 2: "Re-telling 'Painted Skin 2' from the Perspective of Xiao Wei" with an original version; Case 3: "Telling the Story of American Psychopaths from the First Person Perspective of Patrick Beterman" with an onomatopoeic version;

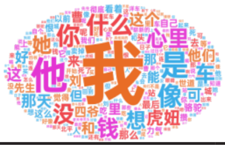
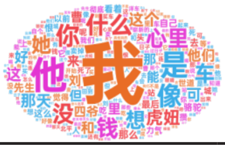
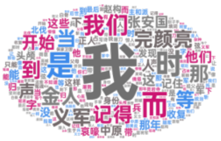
In the first comparison experiment of the paronomasia version and the original version of ‘Painted Skin 2’ from the perspective of ‘Xiao Wei,’ based on the acoustic indicators presented by the samples, the fundamental frequency range of the paronomasia version is 55.98~238.69Hz with a mean fundamental frequency of 165.98Hz, while the original version has a fundamental frequency range of 144.78~351.16Hz with a mean fundamental frequency of 218.83Hz, achieving a restoration ratio of 75.50%. Additionally, both versions show a high degree of overlap in the low, medium, and high frequency ranges below 1000Hz, between 1000~2500Hz, and above 2500Hz. However, it is noteworthy that the sound wave patterns of the paronomasia version exhibit regular changes and pauses, with an average speaking rate of 5.2 words per second, whereas the original version is somewhat irregular, with noticeable long pauses. This suggests that the AI-generated voice quality has reached a level comparable to the original in terms of pitch and sound quality, but its ability to mimic rhythm, emotion, and other emotional aspects still needs improvement.

In the cross-linguistic comparison experiment 2 of ‘Revisiting “Painted Skin 2” from a’ Xiao Wei ‘perspective’ and ‘Telling the Story of American Psychiatric Patients from a’ Patrick Beterman’s’ first-person perspective ‘in their paralinguistic versions, there are significant differences in fundamental frequency and average speaking rate between the two. The maximum fundamental frequency of the English version 3 is 408.41Hz, significantly higher than that of the Chinese version 1, 238.69Hz, while its average speaking rate is only 3.4 words/s, much slower than the Chinese version 1’s average speaking rate of 5.2 words/s. This suggests that AI paralinguistic technology exhibits certain differences or tendencies in language cloning. Although the pitch of the English cloned audio is higher than that of the Chinese cloned audio, and its sound resonance and pauses are highly consistent with the Chinese cloned audio range, the excessively slow average speaking rate inevitably leads to a thin text context and delayed image interpretation, significantly reducing user immersion. This provides an optimization direction for future cross-linguistic AI paralinguistic applications.

Table 4 Generative text analysis

subject investigated		Case 1	Case 2	Case 3
type		film and television	literature	history
Emotional dimension	Subjective word frequency	118	114	34
	Frequency of emotional words	975	539	294
	Emotional type ratio	21.2:46.4:32.4	12.2:51.9:35.9	14.9:58.3:26.8
	Emotional intensity assessment	-0.1364	-0.5797	-0.3173
The pragmatic dimension	Frequency of conjunctions	49	21	11
	Contextual style	pathos	banter	roused
	Subject-verb form	declarative sentence	exclamatory sentence	declarative sentence

Content dimension	Cultural extension	Rituals of worship;
		Strange Tales from a Chinese Studio;
			Beijing rickshaw driver culture; a representative work of Chinese literature	The representative of the Chinese bold school

	Time extends	The complexity of human nature is explored through the theme of surrealism	Expose the social life of the lower class under the rule of the Beiyang warlords in old Beijing	The Southern Song dynasty was invaded by the Jin Dynasty, and the literati could not serve their country
Word cloud				

Note: Case 1: "Revisiting 'Painted Skin 2' from the Perspective of Xiao Wei"; Case 2: "The Night Xiao Fuzi Hanged, Xiangzi Finally 'Died' — Lao She's Most Painful Moment"; Case 3: "The Poem I Am Most Proud of in My Life Was Written with Jin Ren's Brain Bile! Xin Qiji"; The emotional types are ranked in order of positivity, neutrality, and negativity; the emotional intensity is assessed by averaging the sentiment scores of the segmented words;

（2）Generative text dimension: measurement of emotional theme analysis

This section selects the only AI-generated sound video with over one million likes from the film and television category, “Recounting ‘Painted Skin 2’ from Xiao Wei’s Perspective,” the only AI-generated sound video with over ten thousand likes from the literature category, “The Night Xiao Fuzi Hanged: Xiangzi Finally’ Died, ‘” and the only AI-generated sound video with over ten thousand likes from the history category, “My Most Proud Poem in Life Was Written with Jin Ren’s Brain Juice! Xin Qiji.” These videos are used as case studies. Using qualitative coding tools like Nvivo and WeWordCloud, the analysis examines different types of generative dialogue texts from emotional, pragmatic, and content perspectives. The aim is to explore whether the emotional semantics and content depth of generative text can help address the acoustic defects of AI-generated sound technology.

In terms of emotional dimensions, the subjective and emotional words in the AI-generated audio-video text of Case 1 are the most frequent, with 975 and 118 instances respectively, significantly higher than those in Case 2 (a literary AI-generated audio-video) and Case 3 (a historical AI-generated audio-video). However, Case 2 has a higher emotional intensity of-0.5797, while Case 3 has a higher proportion of neutral emotions at 58.3%. In terms of pragmatic dimensions, the frequency of conjunctions in the AI-generated audio-video text of Case 1 is higher than in Case 2 and Case 3, with 49 instances. The main sentence structure in Case 2 is predominantly exclamatory. This suggests that AI-generated audio-video content for films and TV series relies more on emotional and subjective words to enhance narrative integration and emotional transmission due to the need for plot and emotional rendering. In contrast, AI-generated audio-video content for literature tends to use the anger of ordinary people to convey sharp negative emotions due to its focus on human nature and social issues. AI-generated audio-video content for history focuses more on factual statements and biographical introductions, with relatively restrained emotional expression. The coherence and intensity of the plot presented in these texts are directly proportional to their final traffic.

From this perspective, the generative texts of different types of films and TV shows and their AI-generated cloned voices exhibit a significant intertextual effect. On one hand, the cloned and restored voice gives the text an ‘eyewitness account’ feel. On the other hand, the text, which combines emotion and content, provides the voice and characters with an ‘observer’s revelation’ omniscient perspective. These two elements complement each other, weaving another layer of space and life cycle into the plot of films and TV shows. This allows the audience to spontaneously explore deeper social dynamics, cultural changes, and human nature through sensory stimulation.

4.1.3Technical path

Based on the comprehensive application of AI paronomasia technology in first-person film and TV drama narration and the multimodal discourse analysis of samples, this section integrates the narrative-coordination model (N/CM), the operational mechanism of AI paronomasia technology, and the immersive communication paradigm developed by scholars Cummings et al. (2015), Agrawal et al. (2019), and Cao Zhihui et al. (2024). This outlines the following technical path (see Figure 2):

As illustrated in the figure, the technical approach of AI-sound-enabled first-person monologue narration can be roughly divided into three layers: the narrative layer, the effect layer, and the function layer. In the previous section, the effect layer and the function layer were explained through literature review and quantitative analysis. However, the narrative layer, which focuses on how perspective shifts enhance user immersion and emotional perception, was not covered. Therefore, Study 2 will conduct user behavior experiments using perspective shift as a mediator, based on the aforementioned technical roadmap, to explore the narrative advantages of the first-person perspective in immersive communication.

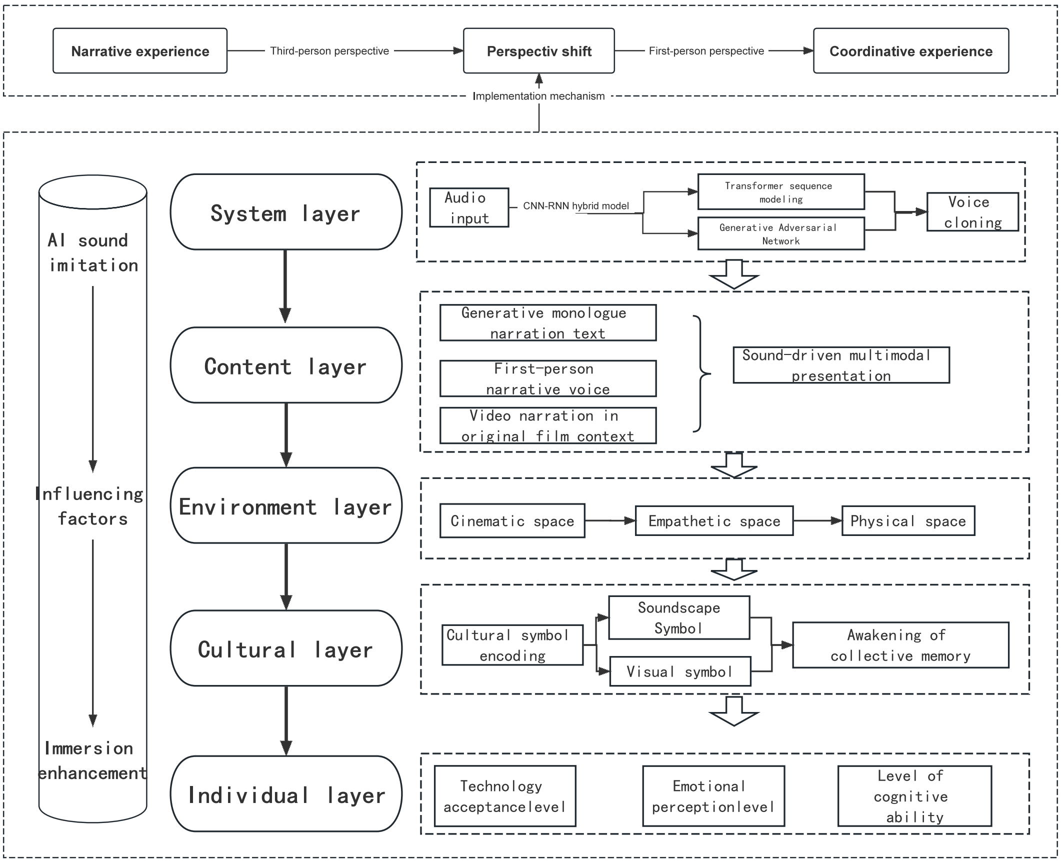


Figure 2. Technical roadmap of AI sound simulation enabling first-person monologue narration

4.2 Study 2: Narrative advantages of AI paronomasia first-person monologue narration from the perspective of inter-group experiments

This study utilized SPSS27.0 and the PROCESS macro plugin developed by Hayes for data analysis. Given that the data were collected through self-reporting by participants and the study involved relationships among multiple variables, it was necessary to conduct exploratory factor analysis and common method bias testing. First, in the questionnaire design phase, the study employed methods such as anonymous surveys, reverse scoring questions, and screening for response time to clean and control the data. Second, before data analysis, Harman› s single-factor test was used to check for common method bias. An unrotated exploratory factor analysis was conducted on all items, and the results indicated that the KMO sample adequacy index of the scale was 0.95> 0.5, the Bartlett› s sphericity test was significant at 0.00 <0.05, and there were two common factors with eigenvalues greater than 1, with the first common factor explaining only 48.7% of the total variance. This ratio is significantly lower than the 50% threshold commonly used in previous studies, indicating that there was no significant common method bias in this study.

Table 5 Validity test table

KMO and Bartlett test		
KMO sample adequacy index.		.951
Bartlett sphericity test	Approximate chi-square	4303.418
	free degree	153
	conspicuousness	.000

4.2.1 Descriptive analysis

This study employed purposive sampling to select participants, primarily from the Beijing and surrounding areas. This was due to two main reasons: first, as an international center for scientific and technological innovation, Beijing has a higher awareness of AI sound technology, making it easier to find suitable participants for this study. Second, it was based on the researchers› personal convenience. The study began with social surveys in Daxing District and Fengtai District of Beijing, and before the formal experiment, participants were pre-tested using the Technology Acceptance Measurement Scale (TAM). Participants were then divided into two groups—Group 1 (n=49, high technology acceptance) and Group 2 (n=59, low technology acceptance)—with a total of 108 participants. The study adopted a mixed experimental design with 2 narration perspectives (first-person, third-person) and 2 technology acceptance levels (high, low). The narration perspective was a within-subject variable, meaning each participant had to watch AI sound commentary videos from both the first-person and third-person perspectives. The technology acceptance level was a between-subject variable, with user immersion as the dependent variable. After the experiment, 210 questionnaires were distributed, and 208 valid responses were collected, including 54 males and 54 females. The average age of the sample was 23.51 years (standard deviation 12.96 years).

Table 6 Descriptive analysis

Experimental group				Gender frequency		Age frequency			
				man	woman	Under 18	18-30 years old	31-50 years old	Over 50
AI imitates the first person	High-tech acceptance sample	count	49	25	24	15	19	11	4
		proportion	23.60%	12%	11.54%	7.20%	9.10%	5.30%	1.90%
	Low technology acceptance sample	count	59	29	30	11	26	10	3
		proportion	24.00%	12%	12%	5.30%	12.50%	4.80%	1.40%
The AI imitates the third person	High technology acceptance sample	count	49	25	24	15	19	11	4
		proportion	24.00%	12%	12%	7.20%	9.10%	5.30%	1.90%
	Low technology acceptance sample	count	59	29	30	11	26	10	3
		proportion	28.40%	13.90%	14.40%	5.30%	12.50%	4.80%	1.40%
amount to		count	208	104	104				

According to the descriptive analysis in Table 6, the gender distribution shows a relatively balanced overall distribution. In the high-tech acceptance sample, males slightly outnumber females; in the low-tech acceptance sample, males are slightly less than females. This suggests that males may have a higher acceptance and willingness to use AI sound simulation technology compared to females. Regarding age distribution, those under 18 and aged 18-30 make up a larger proportion, while those aged 30-50 and over 50 are less common. Additionally, the high-tech acceptance sample among those under 18 is higher, whereas the low-tech acceptance sample among those under 18 is relatively lower. It can be inferred that young people have a higher acceptance and willingness to use AI sound simulation technology.

4.2.2 Main effect test

Correlation analysis is a common method for exploring the mutual influence between two variables. For example, the Pearson correlation coefficient r is used to quantify the relationship between quantitative data. The higher the r value between variables, the stronger their correlation. This study employs the bivariate correlation calculation model in SPSS27.0 to test the two factors influencing user immersion and their hypotheses. The results of the correlation analysis are presented in Table 7:

Table 7 Correlation analysis results among the variables

variable	M average value	SD standard deviation	1	2	3
Narrative perspective	1.52	.501	1		
Technology acceptance	3.25	1.27	-.097	1	
User immersion	3.15	.72	-.947**	.101	1

Note: *p<. 05, **p<. 01, ***p<.001

As shown in Table 7, the correlation index between variables is too high, which may indicate multicollinearity. Therefore, this study needs to further measure the variance inflation factor (VIF) between variables to ensure clear boundaries among variables and prevent model regression distortion. The linear regression analysis results are as follows (see Table 8):

Table 8 Analysis results of linear regression coefficient

model		Unstandardized coefficients		Standardization factor	t	conspicuousness	Collinearity statistics	
		B	Standard error	Beta			tolerance	VIF
1	(constant)	5.209	.070		74.799	.000		
	Narrative perspective:	-1.366	.033	-.946	-41.987	.000	.991	1.009
	Technology acceptance	.006	.013	.010	.440	.660	.991	1.009

A. Dependent variable: user immersion

The regression analysis results in Table 8 show that the VIF values for the narrative perspective (independent variable), technology acceptance (mediating variable 1), and user immersion (dependent variable) are all below the 5 threshold set by previous studies, indicating no significant multicollinearity among these variables. Furthermore, the model has a good fit to the data (R^2 =0.897), with significant regression coefficients. The narrative perspective can predict user immersion (b=-0.946, SE=0.033, p<0.000), partially supporting Hypothesis H1. In the entire research model, the narrative perspective has a stronger correlation with user immersion compared to other influencing variables. The narrative perspective significantly predicts the degree of user immersion, which forms the basis for establishing the variable of technology acceptance. It is worth noting that the predicted results do not fully align with Hypothesis H1, and the narrative perspective of AI parroting technology can be categorized into first-person and third-person perspectives. Therefore, further investigation is needed in the subsequent grouped mediation effect tests to determine which narrative perspective has a more significant positive predictive effect on user immersion.

4.2.3 The overall impact of perspective transformation on user immersion

The total scores and the difference between the post-test and pre-test scores of experimental group 1 (high technology acceptance) and experimental group 2 (low technology acceptance) are shown in Table 9.

Table 9 Descriptive analysis of the total score difference before and after measurement of each group

group	stage	M	SD	D-value
Experimental group 1	before measurement	25.02	0.29	13.76
	aftertest	38.78	0.14	
Experimental group 2	before measurement	24.84	0.24	13.64
	aftertest	38.48	0.23	

Paired samples t-tests were conducted on the pre-test and post-test total scores of Group 1 and Group 2. The results showed that the post-test total score of Group 1 was significantly higher than the pre-test total score (t=13.76, p<0.01), and the post-test total score of Group 2 was also significantly higher than the pre-test total score (t=13.64, p<0.01). This suggests that both the first-person perspective and the third-person perspective significantly enhance user immersion. A one-way ANOVA revealed no significant differences in the post-test total scores of the first-person perspective across different groups (F=0.594, p=0.443). Another one-way ANOVA showed no significant differences in the post-test total score difference between the first-person and third-person perspectives across different groups (F=0.069, p=0.793). As shown in Figure 3, during the intervention, the total scores of Group 1 and Group 2 increased while watching the first-person commentary video, but the slope of Group 1 was not significantly different from that of Group 2. This indicates that controlling for the variable of perspective switching had little impact on user immersion, which is not significant. This finding also supports the conclusions drawn from the main effect test.

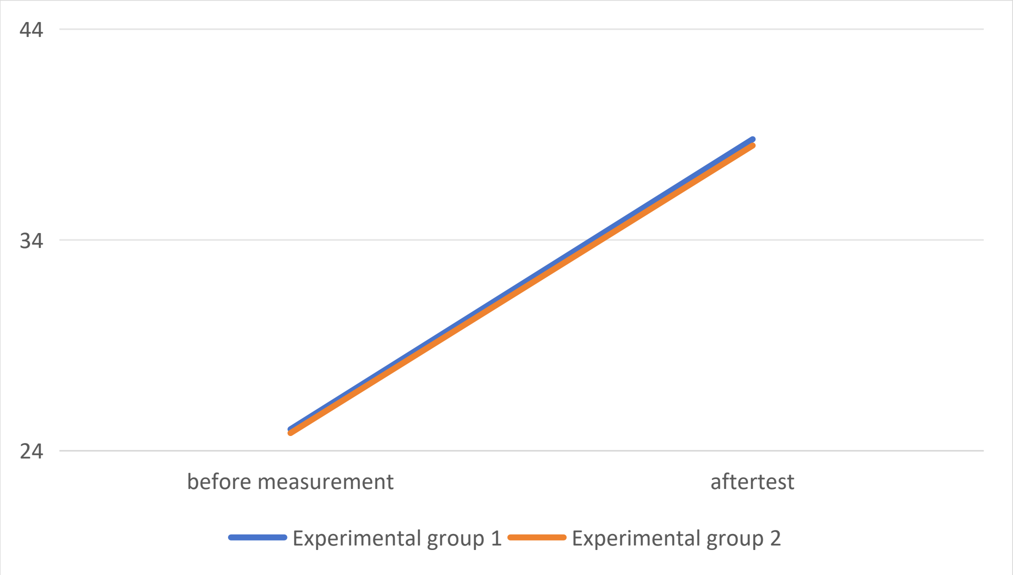


Figure 3 shows the changing trend of total scores before and after measurement in each group

4.4.4Path analysis between variables

This paper employs path analysis to demonstrate the validation results of the relationships among various variables.Based on the descriptive, exploratory, confirmatory factor analysis, and linear regression analyses of the three variables—perspective transformation, user immersion, and technology acceptance—described in the experimental analysis, the significant coefficient relationships are organized into the figure. The path relationships are then corrected and refined based on the original hypothesis model. The resulting path relationship model is presented below (see Figure 3):

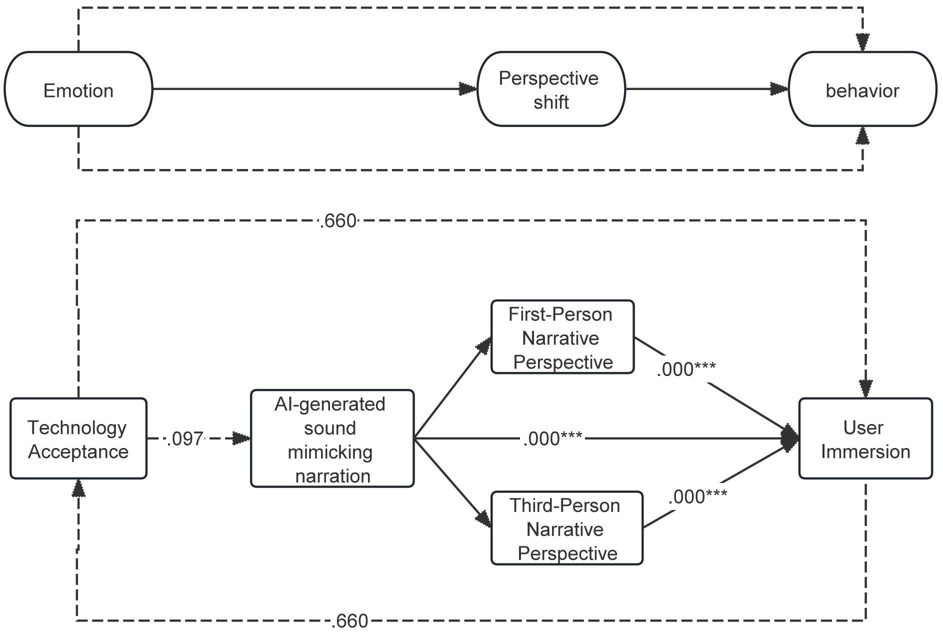

Emotion
Perspective shift
behavior
.660
First-Person Narrative Perspective
.000***
Technology Acceptance
.097
AI-generated sound mimicking narration
.000***
User Immersion
Third-Person Narrative Perspective
.000***
.660

Figure 4 Path model of user immersion by AI sound imitation technology

Based on the coefficient relationships in the path analysis table, the application of AI-sound technology monologue narration significantly positively predicts user immersion. The user's technical acceptance does not significantly affect the degree of immersion or the extent to which users apply AI-sound technology monologue narration, and thus, user immersion cannot positively predict technical acceptance. Combining these findings with the paired samples T-test results from the group experiment, the most significant finding in this study's path analysis is that AI-sound technology monologue narration can positively predict user immersion when moderated by a first-person narrative perspective, and negatively predict user immersion when moderated by a third-person narrative perspective. This result aligns with the mechanisms of hypotheses H1, H2a, and H2b in this study.

5 Insufficient research and measures

While this study, based on the N/CM model, conducted a detailed quantitative analysis of the three variables—narrative perspective (narrative layer), technical acceptance (emotional layer)—of AI paralinguistic technology on user immersion (intentional layer-behavioral), it still has several limitations: (1) The total number of experimental subjects is relatively small, which may lead to a lack of representativeness; (2) The analysis of demographic factors, such as educational background, age, occupation, and place of residence, is limited. These factors could significantly influence user immersion, but presenting data in a purely numerical form does not adequately reflect the user› s immersion experience when using AI paralinguistic technology; (3) The experimental setting is limited, as it was conducted in a specific and relatively monotonous environment, lacking simulation of real-world, diverse usage scenarios; (4) This study primarily focuses on the empirical examination of AI paralinguistic technology in first-person commentary videos, without addressing its broader application areas or copyright issues.

In light of the limitations of previous studies, future research will focus on standardization and interdisciplinary approaches. By designing standardized questionnaire questions and improving sample recovery and cleaning, we aim to enhance the efficiency of quantification and the accuracy of conclusions. We will adopt an interdisciplinary perspective to examine how AI sound simulation technology affects user immersion, aiming to provide theoretical and empirical evidence for enhancing the mechanisms by which AI sound simulation technology influences human emotions and for further refining and standardizing its application. Building on existing data analysis, we can conduct in-depth interviews and group discussions to conduct qualitative research on the emotional states of individual users of AI sound simulation technology, further detailing and supporting the mechanism diagram presented in the text. Additionally, we can expand experimental settings to include various simulated real-world scenarios, collecting immersive experience data from different contexts. Through comparative analysis, we can identify the impact of scene factors on user immersion, making our findings more relevant to practical applications. Given the characteristics of AI sound simulation technology, it is expected to find widespread application in areas such as game sound effects, audiobooks, smart education, and cross-temporal dialogues in museums. Future research can explore these areas further.

Reference

Central People's Government of the People's Republic of China. Notice of the State Council on Issuing the 13th Five-Year National Science and Technology Innovation Plan[N]. (2016-07-28). Retrieved \from https://www.gov.cn/gongbao/content/2016/content_5103134.htm

Central People's Government of the People's Republic of China. Notice of the State Council on Issuing the New Generation Artificial Intelligence Development Plan [N]. (2017-07-08). Retrieved from https://www.gov.cn/gongbao/content/2017/content_5216427.htm

National Film Administration. Notice of the National Film Administration on Issuing the "14th Five-Year Plan for Chinese Cinema"N]. (2021-11-05). Retrieved from https://www.chinafilm.gov.cn/xwzx/ywxx/202111/t20211109_1182.html

Slater, M., & Wilbur, S. (1997). A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. Presence: Teleoperators & Virtual Environments, 6(6), 603-616.

Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire.Presence,7(3), 225-240.

Schubert, T., Friedmann, F., & Regenbrecht, H. (2001). The experience of presence: Factor analytic insights.Presence: Teleoperators & Virtual Environments,10(3), 266-281.

Cummings, J. J., & Bailenson, J. N. (2016). How immersive is enough? A meta-analysis of the effect of immersive technology on user presence. Media psychology, 19(2), 272-309.

Agrawal, S., Simon, A., Bech, S., Bærentsen, K., & Forchhammer, S. (2019). Defining immersion: Literature review and implications for research on immersive audiovisual experiences.Journal of Audio Engineering Society,68(6), 404-417.

Cao Zhihui, Tuo Yanzheng, Han Qiuchen, Chen Ye. The Construction Process and Mechanism of Immersive Experience Scenarios—A Case Study Based on the "Chang'an Twelve Hours" Block[J]. Foreign Economics & Management, 2024, 46(9): 67-88.

Busselle, R., & Bilandzic, H. (2008, May). Emotion and cognition in filmic narrative comprehension and engagement. In annual meeting of the International Communication Association, Montreal.

Zhang Xinyue, & Asano Noriko. (2023). Cognitive Processes of Immersive Experience in Narrative Films. Psychological Review, 66 (2), 215-238.

Denisova, A., & Cairns, P. (2015, April). First person vs. third person perspective in digital games: do player preferences affect immersion?. In Proceedings of the 33rd annual ACM conference on human factors in computing systems (pp. 145-148).

Ministry of Culture and Tourism. Opinions of the Ministry of Culture and Tourism on Promoting the High-Quality Development of Digital Culture Industry[N]. (2020-11-18). Retrieved from https://www.gov.cn/zhengce/zhengceku/2020-11/27/content_5565316.htm

Potter, T., Cvetković, Z., & De Sena, E. (2022). On the relative importance of visual and spatial audio rendering on vr immersion. Frontiers in Signal Processing, 2, 904866.

China Netcasting Services Association. 2024 Micro-Short Drama Industry Ecology Insight Report [N]. (2025-01-01). Retrieved from https://pdf.dfcfw.com/pdf/H3_AP202501161641945784_1.pdf?1737039555000.pdf

Potter, T., Cvetković, Z., & De Sena, E. (2022). On the relative importance of visual and spatial audio rendering on vr immersion. Frontiers in Signal Processing, 2, 904866.

Gideon.AI and Real-Time Special Effects: Revolutionizing Action and Sci-Fi Movies.[N](2023-7-5)https://www.michaelrcronin.com/post/ai-and-real-time-special-effects-revolutionizing-action-and-sci-fi-movies

Li, Y., Yoo, Y., Weill-Duflos, A., & Cooperstock, J. (2021, December). Towards context-aware automatic haptic effect generation for home theatre environments. In Proceedings of the 27th ACM symposium on virtual reality software and technology (pp. 1-11).

Ministry of Culture and Tourism. Report on the Development of Digital Cultural Industries [R]. (2025-01-07). Retrieved from https://www.renrendoc.com/paper/37810635.html

Xia Qingying, Li Zhonggang. Riding the Wind of AI to Break Through Tomorrow's Waves——Media Industry Investment Strategy Report [N]. (2023-12-07). Retrieved from https://pdf.dfcfw.com/pdf/H3_AP202312111613695459_1.pdf

Potter, T., Cvetković, Z., & De Sena, E. (2022). On the relative importance of visual and spatial audio rendering on vr immersion. Frontiers in Signal Processing, 2, 904866.

Linda Greenwood.Presence and Flow: A heuristic framework toheuristic framework to inform theory and design[N](2003-1-18)http://matthewlombard.com/ISPR/Proceedings/ICA2003/Greenwood.ppt

Sanchez-Vives, M. V., & Slater, M. (2004, July). From presence towards consciousness. In 8th Annual Conference for the Scientific Study of Consciousness.

Baños, R. M., Botella, C., Alcañiz, M., Liaño, V., Guerrero, B., & Rey, B. (2004). Immersion and emotion: their impact on the sense of presence. Cyberpsychology & behavior, 7(6), 734-741.

Sohu.com. Can AI Voice Mimicry Software Really Restore Voices 100%? What's the Effect? [N]. (2025-03-13). Retrieved from https://it.sohu.com/a/870328989_121651285

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Patel, A. K., Madnani, H., Tripathi, S., Sharma, P., & Shukla, V. K. (2024, March). Real-Time Voice Cloning: Artificial Intelligence to Clone and Generate Human Voice. In International Conference on Information Technology (pp. 349-364). Singapore: Springer Nature Singapore.

Zhang Delu. Exploration of a Comprehensive Theoretical Framework for Multimodal Discourse Analysis [J]. Chinese Foreign Languages, 2009, 6(01): 24-30.

Tsinghua University. Immersive Experience Questionnaire Survey [N]. (2024-10). Retrieved from https://www.wjx.cn/vm/mbbcmk1.aspx

Cummings, J. J., & Bailenson, J. N. (2016). How immersive is enough? A meta-analysis of the effect of immersive technology on user presence. Media psychology, 19(2), 272-309.

Agrawal, S., Simon, A., Bech, S., Bærentsen, K., & Forchhammer, S. (2019). Defining immersion: Literature review and implications for research on immersive audiovisual experiences. Journal of Audio Engineering Society, 68(6), 404-417.

Cao Zhihui, Tuo Yanzheng, Han Qiuchen, Chen Ye. The Construction Process and Mechanism of Immersive Experience Scenarios—A Case Study Based on the "Chang'an Twelve Hours" Block[J]. Foreign Economics & Management, 2024, 46(9): 67-88.

Podsakoff, P. M .Self-Reports in Organizational Research: Problems and Prospects[J].Journal of Management, 2016, 12(4):531-544.DOI:10.1177/014920638601200408.

Belsley, Kuh, Welsch. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity [M]. 1980.