frontiers | Frontiers in Virtual Reality

TYPE Original Research
PUBLISHED 14 November 2025
DOI 10.3389/frvir.2025.1691405

OPEN ACCESS

EDITED BY
Nelly Lagos San Martín,
University of the Bío Bío, Chile

REVIEWED BY
Mark Grimshaw-Aagaard,
The Hong Kong University of Science and Technology (GZ), China
Tom Garner,
Sheffield Hallam University, United Kingdom

*CORRESPONDENCE
Zhenzhen Li,
zhenzhen.li@bristol.ac.uk;
leew55437@gmail.com

RECEIVED 28 August 2025
REVISED 18 October 2025
ACCEPTED 31 October 2025
PUBLISHED 14 November 2025

CITATION
Huang Q and Li Z (2025) Examining immersive audio's role in flow and gaze behavior in pseudo-VR environments.
Front. Virtual Real. 6:1691405.
doi: 10.3389/frvir.2025.1691405

Examining immersive audio's role in flow and gaze behavior in pseudo-VR environments

Quncan Huang[1] and Zhenzhen Li[2]*

[1]The University of Sheffield, Sheffield, United Kingdom, [2]University of Bristol, Bristol, United Kingdom

The audience's sense of presence diminishes significantly when 360° videos are presented on a flat 2D screen. This study investigates whether immersive audio can compensate for the loss of visual immersion, enhancing viewers' concentration and emotional engagement on non-VR devices. A key question is thus raised: when production resources are limited, should creators prioritize auditory or visual enhancement to improve the overall user experience? In order to answer this, we had 150 subjects watch several different types of video, some VR-like and others not, with sound either visualized on the ceiling or standard audio. Where possible, we monitored eye movement during viewing, and emotional involvement was assessed using a simple questionnaire afterwards. Results speak for themselves since implementing auditory improvement has been seen to increase involvement or concentration as well. In cases where only VR videos are provided, concentration and interest are not easily achieved. Therefore, we can conclude that in settings outside head-mounted devices, auditory qualities vastly outweigh visual enhancement if immersion is desired. This paper presents a simpler and more practical solution to the problem for creators of VR industrial material. It has been shown that even if using VR devices proves impractical or too costly, the population can greatly enhance the emotional effects of watching VR videos by focusing more on auditory technology. This presents a new mode of thought for producing products, which may enhance the solutions derived from wide usage and decrease the expense necessary for the advantages of an immersion product.

KEYWORDS

flow theory, eye-tracking, emotional and physiological response, immersive audio, film

Introduction

The immersive experience created by modern movie technology, on the other hand, has the superficial effect of inducing a "dreamlike" state in the audience, through multisensory software and hardware, real-time and controllable, which allows for its fallacious apprehension; its real aim, nevertheless, may lead to a "flow-trap" phenomenon, where prolonged sensory engagement can result in passive absorption rather than active enjoyment (Yu and Lo, 2023). Shaw gives man a resort, and the spectator easily becomes a passive prey to a condition of concentration or absorption in which he thoroughly loses track of time, in which increased enjoyment is experienced. In this state, a condition of enjoyment replaces the speculation. Csikszentmihalyi (2014) describes this state as the "fusion of action and awareness" and the "dissolution of self-consciousness." Additionally, flow depends on individual traits, including curiosity and preferences (Parvizi-Wayne et al., 2024). However, technological standardization presents a universal "immersion template" that ignores cognitive differences among individuals

(Howlett and Paulus, 2017). What is more serious, however, is the issue of visual bias: VR developers enhance the field of view and resolution to strengthen the illusion of reality with 3D, while audio takes a back seat. The inevitable result is a dilemma: when VR content is viewed via a 2D render that was painstakingly immersive in visual design, this immersion is dramatically reduced. While immersion in this study refers to a technologically induced sensory engagement with audiovisual media, flow denotes a psychological state of deep absorption characterized by concentration and loss of self-awareness. This clarification helps distinguish perceptual immersion from the cognitive and emotional processes defined by Flow theory. Consequently, the research focuses on a practical challenge: how to provide a greater immersive experience of the VR product than is possible in cases where complete immersion is not feasible. In this connection, the research probes the influence audio has on emotions. For our purposes, we investigate whether the uptake of emotions and cognition can be increased through the use of immersive audio, where complete immersion in the specific experience is not possible. We compare immersive and conventional audio effects in triggering the same emotions to challenge the view that new media will inevitably replace old media (Lehman-Wilzig and Cohen-Avigdor, 2004). For clarification, the basic concepts of the study have been formulated as follows: A "VR video" is a video segment that utilizes 360° panoramic rendering, which is seen as a display on a 2D screen. An "ordinary video" is a video segment with a fixed angle of view that is also presented on a 2D display. "The psychological anticipation of the immersive label" explores how subjects anticipate the experience based on the researchers informing them about the kind of auditory experience they will have. The preference for a specific stimulus pattern" indicates participants' systematic emotional preference for certain audio-visual combinations, showing clear interest and heightened attention. This study aims to investigate three possible underlying mechanisms. First, determining whether the differences between the two arise from the technical implementation of video types—virtual reality videos versus regular videos. Second, participants' psychological expectations of the "immersive" label, and discomfort caused by mismatched audio and dynamic visual cues. Third, there is a systematic preference in the evaluation for specific stimulus patterns (Husselman et al., 2024). An explanation of eye gaze abnormality deviations caused by immersive audio in persons with certain cognitive characteristics in pseudo-VR defects regarding extreme outliers in eye-tracking data can show whether persons still derive added value from "technological commitment (immersive audio)" and "technological reality (VR content video)" in which Chion's synchronization theory is broken (Chion, 2019). It is indicated that to a great extent, technological development depends on technical integrity, cognitive congruence of user and technology, and verisimilitude of the media (Shang et al., 2025). Secondly, we noted the importance of immersive audio as a characteristic of sophisticated technology. Nevertheless, the state of psychological presence depends on the relationship between the user's cognitive processes and the environment provided by the media, rather than on any single variable (Ananthabhotla et al., 2021). The label "immersion audio" without confusion implies a promise of technical excellence. However, failing to provide the key environmental factors necessary to elicit psychological presence creates a significant gap between expectation and reality, which erodes participants' experience. Hence, this study investigates whether immersive audio can evoke more "automatic" and less regulatable emotional responses in VR content videos. Its potential value lies in demonstrating a more inclusive and efficient path for optimizing VR experiences, namely ensuring the integrity of auditory immersion and visual content.

Theoretical background

Emotional resonance and audio

As the rapid development of virtual reality technology, the sense of presence of VR experience not only originates from vision, but is also related to the mental mechanism of hearing. Inhibition of the μ rhythm in the human auditory system is greater in virtual reality than in conventional video, indicating that immersive audio is more effective in stimulating the nervous system and promoting empathy (Gallese, 2005). There is evidence suggesting that in virtual reality environments with immersive audio, the inhibitory effect of the μ rhythm is more significant than when watching ordinary videos. This neurological principle has long been applied in art and media practice. As the sound designer of "Apocalypse Now", Walter Murch said, film music guides the audience into a place where they can endlessly imagine, creating imagined sounds from non-existent ones (Kujawska-Lis, 2000). Music evokes the audience's emotions and psychology, inspiring their imagination and creativity (Stock, 2009). Virtual Reality combines immersive sound, visuals, and interactive components in a 3D virtual environment to foster an intense sense of presence that tricks the senses and impacts emotions (Somarathna et al., 2022). Similarly, "Avatar 2: Waterways" aims to create a comparable immersive experience on the screen through advanced projection and sound technologies. As scholars have pointed out, this film represents a significant advancement in film technology (Li and Xie, 2023). Specifically, director James Cameron utilized laser projection for more realistic light sources and Dolby Atmos to simulate environmental sounds, including insects, wind, creatures, water currents, and collisions. The work of these elements, similar to a VR system, enhances the audience's psychological presence and the feeling of weightlessness in the sea and forest.

Flow theory and pseudo-VR

This study is based on the theoretical framework proposed by Mihaly Csikszentmihalyi in 1975, namely the "Flow Theory" and the "Three-Channel Model". The aim is to examine the impact of different audio combinations on users' emotions and cognition. In this study, a structured viewing format was used—non-immersive 2D screen playback—to give participants a clear objective (viewing stimuli), fulfilling the key prerequisite for entering the flow state.

Building on the mechanisms of Flow theory—like heightened attention (Hypothesis 1), reduced self-awareness (Hypothesis 2), variation stemming from cognitive load (Hypothesis 3), and feedback-induced anomalies (Hypothesis 4)—we formulated the following hypotheses:

Hypothesis 1: Drawing from the "focus and immersion" dimensions of Flow theory, participants will exhibit longer gaze duration during video clips accompanied by immersive audio compared to those with regular audio.

Hypothesis 2: The subjective ratings and gaze durations of virtual reality content videos are higher than those of ordinary videos, likely due to the loss of self-awareness.

Hypothesis 3: Based on feedback and challenge-skill balance, immersive audio can elicit more positive subjective ratings (e.g., spatial quality, preference) than conventional audio. However, there are extreme values due to individual differences in cognitive load.

Hypothesis 4: When the complexity of the technology exceeds the cognitive load of the participants, the cognitive redistribution mechanism will be triggered, resulting in extremely abnormal values far higher than those in the virtual reality content videos. Csikszentmihalyi stated that when skills and challenges are in balance, a state of flow occurs. This state closely links our abilities with our perception of the challenge. In this study, our immersive audio-visual design created the necessary conditions for flow, such as balancing skills and challenges, providing immediate feedback, establishing clear goals, promoting concentration and focus, and integrating action awareness (Csikszentmihalyi, 1975). Firstly, the state of flow depends on clear goals and immediate feedback. In the experiment, we explicitly gave the participants the instruction "watch the video and answer the questions", which gave them a sense of participation. When the participants were performing the task, the immersive audio (binaural channels) of the spatial audio rendering technology provided them with precise information about the direction, distance and environment of the sound source, offering clear stereoscopic feedback that helped the participants quickly perceive changes and respond (Webster et al., 1993). This confirms Hypothesis 3, which expects that increased immediate feedback will improve the sense of presence from immersive audio. The video of the virtual reality content was shown on a two-dimensional screen, but its high resolution and large view improved visual details and further increased the immediacy of feedback. Furthermore, adjusting the complexity of the audio-visual stimuli fulfills the "challenge" to cognition. The mismatch between the relatively simple spatial visual information in the pseudo-virtual reality videos and the multi-dimensional spatial audio increases the cognitive load and the difficulty of perceptual processing. When the technical challenges match the skills, the participants can effectively process multi-sensory information and enter a deep engagement or smooth state. This is reflected in the positive subjective evaluations (Hypothesis 2, Hypothesis 3) and efficient eye movement patterns (Hypothesis 1, Hypothesis 2). However, when the technical complexity of the pseudo-virtual reality videos combined with immersive audio exceeds the individual's cognitive capacity, the cognitive resource reallocation mechanism is triggered (Engeser and Rheinberg, 2008). This overload may lead to anxiety rather than a smooth feeling, and is ultimately manifested in the high differences or opposite extreme situations in Hypothesis 3 and Hypothesis 4. The key characteristics of the flow state include high concentration and the loss of self-awareness (Ulrich et al., 2014). The longer the gaze duration in the experiment, the deeper the information processing, the better the attention maintenance, and the stronger the anti-interference ability. For example, compared to ordinary videos, pseudo-virtual reality videos have longer gaze durations due to their richer information content, which supports hypotheses Hypothesis 1 and Hypothesis 2. Even when merely passively viewing, highly consistent multi-sensory input can enhance the "presence", which is a low-level internal interaction form achieved through eye movements (Swann et al., 2012). To capture this state, we evaluated the participants' concentration, pleasure, and perception of time changes through a subjective questionnaire.

Methods

Procedure

First, we recruited volunteers from social media platforms, such as Xiaohongshu, and the residential area in Houhu, Wuhan. Volunteers had to be at least 18 years old, have normal or corrected vision, and never have experienced dizziness or severe symptoms from 3D or VR content video. Participants were not pre-screened for hearing acuity, which may introduce minor individual variability in the perception of spatial audio. This limitation is addressed in the discussion and future research section. Online volunteers understood the experiment via WeChat and completed it independently, while offline volunteers did so with our help. After watching videos, all completed a questionnaire in Supplementary Material S1. The participants received a monetary reward of 10 yuan. The study involved no medical procedures, sensitive data, or potential risk, so according to institutional guidelines, no ethics review was required. The stimuli were presented to each participant in the same order, predetermined in the experimental designs, ensuring that the video and audio types followed a logical, nonrandomized sequence. It should be clearly noted that in this experiment, due to the consistent sequence of stimulus viewing, it is impossible to separate the sequence effect from the main effect of the experiment itself.

Questionnaire

This study focuses solely on the questionnaire related to eye-tracking, which is central to the Flow theory (Supplementary Material S2). Among the questions, Q1 (involuntary musical imagery) directly measures the "immediate feedback" of the Flow theory, Q2 (total focus on the video content) assesses "concentration," Q3 (immersion) gauges "disappearance of self-awareness," and Q4 (excitement) evaluates "intrinsic reward" (Jackson and Marsh, 1996). Although Q5 (physiological response) does not align with the core dimensions of the Flow theory, this study considers it an auxiliary indicator to explore the potential link between psychological experience and physiological arousal (Jackson and Marsh, 1996).

Stimuli

The stimuli included various video and audio combinations: ordinary video with regular audio (OV + RA), immersive audio (OV

+ IA), VR content video with regular audio (VR + RA), and immersive audio (VR + IA). VR visuals here refer to VR game excerpts displayed on a screen, rather than through a head-mounted display (HMD) in an interactive VR environment. Thus, the VR condition is better described as VR content video. The ordinary video featured a 40-s underwater segment from "Avatar 2: Waterways" with the original soundtrack (IA) or an audio tagged "Sea" (RA). The VR content video was also 40 s, from "Attack on Titan," with the original soundtrack (IA) or an audio tagged "Battle" (RA). Immersive audio refers to the use of dual-ear rendering technology to precisely position each sound element in a 3D sound field, creating an immersive auditory experience. Ordinary audio uses a mono audio format, thus lacking spatialization information and a stereoscopic sound field. The "scene content matching" strategy is employed to ensure that the differences between the two types of audio are in the presentation of the spatial sound field, that is, creating independent audio tracks for the same scene with highly consistent themes, durations, and emotional tones. This strategy enables us to control the content variables and facilitate the examination of the reasons for the differences in the audio-induced experiences.

Apparatus

All videos are played through the Gorilla platform. Besides Firefox (which cannot record eye-tracking data), other browsers like Chrome and Microsoft Edge can be used. The audio samples are at 44.1 kHz, 16-bit, stored as uncompressed WAV, and are in mono plus spatial metadata format (MPEG-H 3D Audio standard). Immersive audio uses hardware that supports Dolby Atmos rendering (Dolby sound headphones) to render dual-channel audio in real time, simulating the precise positioning of sounds in a three-dimensional space. Regular audio is standard 2.0-channel stereo, which is played back through the same pair of headphones. During playback, it must be rendered into a binaural sound field via headphones supporting head tracking. Online participants need headphones that produce a 3d spatial sound, while offline participants use our Dolby Atmos Headphones (like AirPods Max) for accurate sound field reconstruction based on head movements.

Sample

We recruited 219 adults; 17 did not finish, 52 had invalid lie detector results, so 150 were analyzed (76 females, 74 males, mean age = 33, S.D. = 11.36, range = 18-60). While online and offline numbers were not recorded, demographic info was complete. Volunteers were grouped by age: ≤35 years (N = 102) and >35 years (N = 48), with 35 years chosen as a cutoff due to findings in auditory physiology, neuroscience, and psychology. These studies show that the auditory cortex is plastic, vestibular function declines naturally, and cognitive load shifts during adolescence (up to age 35) (Eggermont, 2019). Ninety-eight percent of volunteers were from China, so the educational standards of mainland China were adopted (Supplementary Material S3). The average education was 14.49 years (SD = 2.95, range = 6–23 years), with 70.1% having a senior college degree or higher (≥15 years).

Measures

The datasets generated and analyzed during the current study are available in the Open Science Framework repository, Examining Immersive Audio's Role in Flow and Gaze Behavior in Pseudo-VR Environments, under the following URL: https://osf.io/54mke. This experiment used Gorilla's eye-tracking module with Web Gazer to detect faces and predict gaze location, using 5-point calibration and drift correction before each session. The recording time and percentage spent were displayed. The total gaze duration is the fundamental measure of eye movement tracking for this experiment. It refers to the total amount of time (measured in milliseconds) the subjects' eyes are fixated upon the viewing area for the video playback, which indicates the degree of attention-in-being. All analyses were completed in IBM SPSS Statistics 27.0. The gaze times for the four areas of the variables (VR, OV, IA, RA) were summed separately. Comparing the mean differences in gaze time between VR content video and OV, as well as between IA and RA, by a *Within-Subjects Design*, which improves sensitivity to task effects by eliminating interference from individual differences. Both VR content video and OV data were collected from the same sample of volunteers (N = 150), ensuring direct comparability and verifying the following hypotheses:

1. The difference values (VR-OV) did not meet normality ($W = [0.29]$, $p < 0.001$), so a Wilcoxon signed-rank test was used.
2. No outliers requiring exclusion (exceeding mean±3SD) were found. Despite high variability, we report two-tailed test results with effect size (Cohen's d) and 95% CI to prevent Type II errors. The significance level is $\alpha = 0.05$ (two-tailed), and the calculation formula is $d = (VR\text{-}OV)/SD_{pooled}$, where $d = 0.08$ is insignificant. Due to outliers, the Wilcoxon signed-rank test was used to assess median gaze differences between IA and RA, VR content video, and OV.

The scale in the questionnaire comprises one dimension (immersion and psychophysiological responses) with five items, rated on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree). The dimension score reflects the combined item scores, with a total score ranging from 1 to 5 points. Reliability and Factor analysis were used to calculate α and KMO values. One-way ANOVA tested effects of conditions: ordinary video + regular audio, ordinary video + immersive audio, VR content video + regular audio, VR content video + immersive audio (OV + RA, OV + IA, VR + RA, VR + IA) on variables (Q1-Q5). The Levene test revealed that all dependent variables violated the assumption of homogeneity of variance ($p < 0.05$). A robust Welch ANOVA was used to replace traditional ANOVA in testing mean differences. Results showed significant differences ($p < 0.05$), and ANOVA was also significant ($p < 0.01$). Pairwise comparisons used the Games-Howell *post hoc* test, with effect sizes calculated using η^2, at $\alpha = 0.05$.

TABLE 1 Comparison of gaze duration under Immersive audio (IA) and Regular audio (RA) (N = 150).

Condition	*M*(sec)	SD (sec)	Median	Mean diff	t(df)	p	d	95% CI
IA	46,120.03	15,690.93	44.053,43					
RA	42,769.85	10,152.75	40,767.45					
Comparison		18,028.17		3,350.18	2.28 (149)	0.024*	0.186	[441.50, 6,258.86]

IA, immersive audio; RA, regular audio; M, mean; SD, standard deviation; Mean diff, Mean difference (IA-RA); df, Degrees of Freedom; CI, Confidence Interval. The unit of fixation time is seconds. *Indicates $p < 0.05$. Data are median (IQR) since the Shapiro-Wilk test indicated non-normal distribution ($W = .52$, $p < .001$). The Wilcoxon signed-rank test was used to compare IA, and RA.

TABLE 2 The results of the wilcoxon signed-rank test (N = 150).

Ranks				Test statistics[d]		
		N	Mean rank	Z	Sig. (2-tailed)	r
IA - RA	Negative ranks	30[a]	87.67	−5.690[e]	0.000	0.46
	Positive ranks	120[b]	72.46			
	Ties	0[c]				
	Total	150				

***$p < .001$; gaze duration in ms.
[a]IA < RA
[b]IA > RA
[c]IA, RA.
[d]Wilcoxon Signed Ranks Test
[e]Based on negative ranks. Effect size is calculated as r = |Z |/n per Fritz et al. (2012). r = [Value] indicates effect size.

Result

Reliability, validity, and welch

The Cronbach's alpha coefficient was computed in order to assess the internal consistency reliability of the scale. The coefficient yielded $\alpha = 0.872$, indicating good reliability of the measure. The KMO was found to be 0.868, indicating good adequacy of sampling, and Bartlett's test was found significant ($\chi^2 = 1411.527$, df = 10, $p < 0.001$), suggesting that the data were suitable for factor analysis. The exploratory factor analysis showed a single factor with an eigenvalue greater than 1, accounting for 66% of the variance in the scale. All five items had factor loadings greater than 0.6 (Range = 0.772 - 0.870), hence supporting the validity of the scale. Levene's test showed there were no significant main effects on the dependent variables ($F = 1.868$ - 17.574, $p < 0.05$), inducing a violation of homogeneity of variances in ANOVA. Subsequent analysis using Welch's procedure showed significant main effects ($F = 77.542$ - 111.725, $ps < 0.001$, $\omega^2 > 0.2$), so a statistically significant effect.

The general advantages of IA

Table 1 shows *descriptive statistics and t-test* results comparing immersive audio (IA) and regular audio (RA). The IA time (M = 46,120 ms, SD = 15,6901, Mdn = 44,053.43) exceeded RA time (M = 42,770 ms, SD = 10,153, Mdn = 40,767.45), with a median difference of 3,285.98. Gaze times in IA and RA were positively correlated (r = 0.076), but not statistically significant (p = 0.354). A significant mean difference between IA and RA showed by The Paired-Sample t-test ($\Delta M = 3,350$, 95% CI [441.50, 6,258.86], $t(149) = 2.28$, $p < 0.05$). The small effect size (d = 0.186) indicates limited practical significance. The wide confidence interval and large SD diff (18,028.17) show high variability in responses to the interaction patterns.

Although volunteers spent more time looking at IA, the Wilcoxon test revealed a different pattern (Table 2). RA had a higher mean rank (87.67) than IA (72.46, $Z = 5.690^b$, $p < 0.001$). The number of "RA > IA" cases was smaller (30), but the number of "IA > RA" cases was larger (120). Extreme outliers in RA resulted in high ranks, which led to the rank test indicating that RA's average rank was higher due to these rare, extreme negative values (Figure 1).

Table 3 shows the Game-Howell *post hoc* test for dependent variables (Q1-Q5). When audio type was IA, VR + IA scores exceeded ordinary video + immersive audio (OV + IA) (M diff = 1.067-1.124, ds ≥ 0.8, ps < 0.001, 95%CI [0.58,1.48]). When the audio type was RA, VR + RA scored higher than OV + RA (M diff = 0.493-0.967, ds ≥ 0.8, ps < 0.001, 95%CI [0.51,1.46]). When the video type was VR content video, VR + IA scored higher than VR + RA (M diff = 0.840-0.947, ds ≥ 0.8, ps < 0.001, 95%CI [0.6, 1.18]). When the video type was OV, OV + IA scores exceeded OV + RA (M diff = 0.493-0.967, ds = 0.48-1.01, ps < 0.001, 95%CI [0.19,1.25]).

Slight differences in video types

Table 4 presents descriptive statistics for the experimental conditions on variables Q1-Q5. Scores in VR + IA and VR + RA conditions in Q1-Q5 were higher than OV + IA and OV + RA, with VR + IA highest on Q2 (M = 4.3, SD = 0.817) and OV + RA lowest on Q5 (M = 2.1, SD = 1.054).

Table 5 shows *Descriptive Statistics and Paired-Sample t-test* results for VR content video and OV comparison. The mean for VR

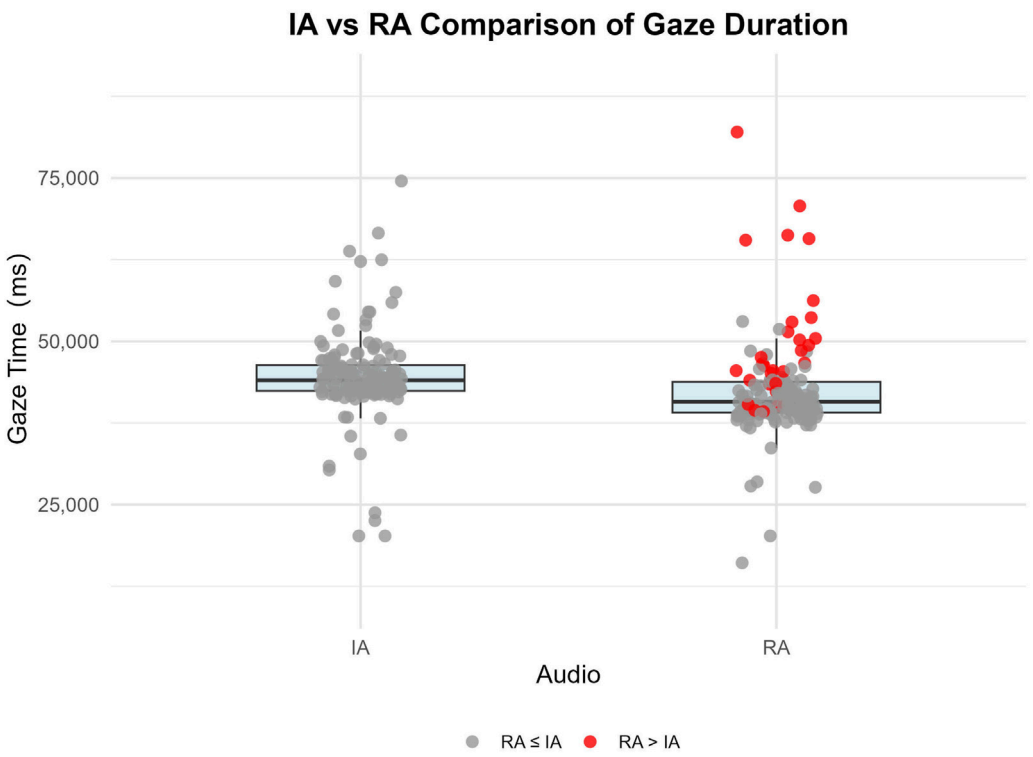

FIGURE 1
Gaze duration comparison: Immersive Audio (IA) vs. Regular Audio (RA). Red dots mark individuals more attracted to normal audio anomalies (RA's gaze duration was longer than IA).

TABLE 3 Post-hoc key pair comparisons (Games-Howell) (N = 150).

DV1	IA	VR	RA	OV
DV2	VR > OV	IA > RA	VR > OV	IA > RA
		M(p)		
Q1	1.240***(0.001)	0.893***(0.001)	0.967***(0.001)	0.620***(0.001)
Q2	1.100***(0.001)	0.847***(0.001)	0.947***(0.001)	0.693***(0.001)
Q3	1.173***(0.001)	0.840***(0.001)	0.827***(0.001)	0.493***(0.001)
Q4	1.067***(0.001)	0.947***(0.001)	0.840***(0.001)	0.847***(0.001)
Q5	1.107***(0.001)	0.907***(0.001)	0.493***(0.001)	0.967***(0.001)

DV, dependent variable; VR = VR, content video; M(p), Mean difference (p value), with mean difference rounded to 2 decimal places. Significance: *p < .05, **p < 01, ***p < .001. Full results (including SE, and CI) in Supplementary Material S4. Descriptive stats in Supplementary Material S5.

TABLE 4 Descriptive statistics (*M* ± SD) of the dependent variables (Q1-Q5) and condition (N = 150).

Condition	Q1	Q2	Q3	Q4	Q5
VR + IA	4.07 ± 0.836	4.30 ± 0.817	4.26 ± 0.893	4.21 ± 0.894	4.17 ± 0.918
VR + RA	3.18 ± 0.751	3.45 ± 0.856	3.42 ± 0.964	3.27 ± 0.841	3.27 ± 0.902
OV + IA	2.83 ± 0.781	3.20 ± 0.827	3.09 ± 0.919	3.15 ± 0.937	3.07 ± 0.849
OV + RA	2.21 ± 1.097	2.51 ± 1.008	2.59 ± 1.124	2.30 ± 0.954	2.10 ± 1.054

M, Mean; SD, standard deviation.

TABLE 5 Comparison of gaze duration under Virtual reality (VR) and Ordinary video (OV) (N = 150).

Condition	*M*(sec)	SD (sec)	Median	Mean diff	t(df)	p	Cohen's d	95% CI
VR	45,157.79	9,236.69	43,494.88					
OV	43,732.10	16,233.89	40,816.90					
Comparison		18,004.42		1,425.69	0.97 (149)	0.334	0.079	[-1,479.16, 4330.53]

VR, virtual reality content video; OV, ordinary video; M = mean; SD, standard Deviation; Mean diff, Mean difference (VR-OV); df = degrees of Freedom; Ci = confidence Interval. Fixation time is in seconds. Data are presented as median (IQR), as the Shapiro-Wilk test indicated non-normality (W = 0.49, p < .001). The Wilcoxon signed-rank test was used to compare the VR, content video and OV; groups.

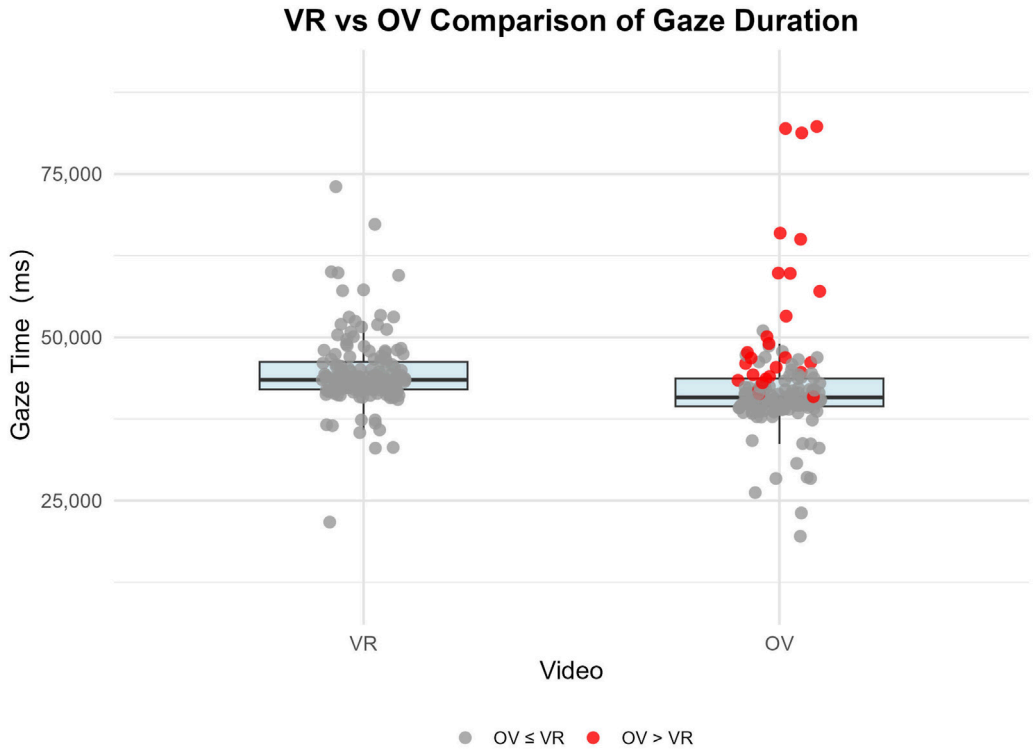

FIGURE 2
Compares gaze duration between VR content video and Ordinary video (OV). Red dots indicate individuals who are more attracted to OV's audio anomalies, exhibiting longer gaze durations compared to those in the VR content video.

content video (M = 45,157.79) was slightly higher than OV (M = 43,732.09), with a median difference of 2,678.08. There is a positive correlation between VR content video and OV gaze time (r = 0.082), but it is not statistically significant (p = 0.316). A Paired Sample t-test showed no significant difference between VR content video and OV (ΔM = 1,425.69, 95% CI [-1,479.16, 4,330.53], t (149) = 0.97, p = 0.334). These results suggest that there are no significant differences or linear relationships between VR content videos and OV. OV (SD = 16,233.89) is significantly higher than VR content video (SD = 9,236.69), indicating the presence of outliers under OV with large values (Figure 2).

Although more volunteers gazed longer at VR content video, the Wilcoxon test revealed a different pattern. The results showed that the mean rank of Ordinary video (OV) (mean rank = 86.10) was higher than that of VR content video (mean rank = 72.85) ($Z = -5.778^{b}$, $p < 0.001$). The results are consistent with the high variability of OV (SD_OV > SD_VR). Hence, although there are fewer cases where OV > VR content video (30) than cases where VR content video > OV (120), the extreme outliers in OV are significantly larger than those in VR content video, resulting in extremely high ranks (Table 6). Moreover, based on the rank test, these few negative values dominate the results, causing the average rank of OV to be higher than that of VR content video.

Discussion

Our multimodal system demonstrated how different audio-visual combos influence gaze duration and the degree of presence experienced by the participants. The study has shown a complex reciprocal relationship between gaze length and subjective ratings, which emphasizes the important role of IA. By examining the definition of flow theory and the technological terms of immersion, an overlap in definition was discovered, which may

TABLE 6 The results of the wilcoxon signed-rank test (N = 150).

Ranks				Test statistics[a]		
		N	Mean rank	Z	Sig. (2-tailed)	r
VR-OV	Negative ranks	30[a]	86.10	−5.778 [b]	0.000	0.47
	Positive ranks	120[b]	72.85			
	Ties	0[c]				
	Total	150				

VR = VR, content video, ***p < .001, gaze duration in ms
[a]VR < OV
[b]VR > OV
[c]VR = OV
[d]Wilcoxon Signed Ranks Test
[e]Based on negative ranks. Effect size r = |Z|/n, as per Fritz et al. (2012). r = [Value] indicates a [small/middle/large] effect.

lead to confusion, for although having a deep experience is one of the things that has been advanced as a requirement of flow state (Csikszentmihalyi), it differs from flow theory. The term immersion in this research is intended to convey an all-inclusive experience through sensory stimulation. This technological immersion is designed to concentrate attention on the virtual environment and elicit a subjective sense of presence. Conversely, flow is a broader and more advanced concept, necessitating a deep self-forgetfulness, characterized by "the dissolution of self-awareness," alongside clear goals, immediate feedback, and an optimal balance between challenge and skill, to attain the optimal experience. In summary, deep flow involves cognitive absorption, but not all immersive experiences can reach the flow state. The Q3 (immersion) specifically highlights the key feature of "measuring the dissolution of self-awareness."

Apparent influence of audio on psychological immersion

The five dimensions—involuntary musical imagery, concentration, cognitive absorption, excitement, and physiological response—measured the Flow experience (Jackson and Marsh, 1996). Welch-ANOVA and Games-Howell tests revealed that four audio-visual combinations had a significant effect on participants' sense of presence. Although the audio is mono in the file, on compatible devices, it provides a significantly enhanced immersive experience compared to regular audio. Significant differences were observed across all five indicators (Q1-Q5) ($p < 0.05$), and the effect sizes were all practically meaningful, indicating that the different media and interaction modes have differential regulatory effects on the core psychological experiences (concentration, cognitive absorption, intrinsic motivation) of Flow. The VR + IA combo was the most effective, outperforming others in all areas (music persistence Q1, cognitive absorption Q2, focus Q3, excitement Q4, physiological response Q5) ($p < 0.05$). This aligns with Csikszentmihalyi's Flow theory, suggesting that immersive VR and interaction design can optimize the "challenge-skill balance" (Csikszentmihalyi, 2014). In contrast, active interaction (IA) enhances "action-awareness integration" with instant feedback (Csikszentmihalyi, 2014). As Romero-Fresco and Fryer (2013) noted, audio in films is information that audiences must process. It provides additional, interconnected information through a "top-down" approach, autonomously engaging audiences' attention and emotions, creating a subjective understanding of the narrative (Romero-Fresco and Fryer, 2013). VR content video only has an advantage over ordinary video when combined with IA (VR + IA > OV + IA). There is no significant difference when combined with RA (VR + RA vs. OV + RA). IA enhances spatio-temporal binding through salient audio information. The greater vistas and vertiginous video presentations of virtual reality movies along with IA engender a greater sense of reality (Q2) and physiological arousal (Q5) (Chion, 2019). When not distinguishing between video types, the amount of gazing time is greater on average by IA (Invincible audio) (M_IA = 46,120.03, M_RA = 42,769.85), and the median comparisons do not reveal any significant differences. Paired t-tests indicate significance ($p < 0.05$), whereas Wilcoxon tests reveal complexity. With these unexpected results came a vast Flow, too deep for description, a "disturbed sense of time," while one as well experienced the "dissolution of self-awareness." This agrees with Sherry's "Flow" of media, as it is possible that a few similar receiving stimuli may enjoy and scheme easily upon attentiveness (Sherry, 2004). Although VR > OV (N = 120) and IA > RA (N = 120) had larger groups, the larger group's average rank was lower than the smaller groups'. Both Wilcoxon test p-values were significant.

Objectively observing the contradiction and unity of time and subjective emotions

This research found that, without differentiating between audio types, the average gaze time for VR content video was higher than for ordinary video (OV) (M_VR = 45,157.79, M_OV = 43,732.10). This matches subjective immersion scores, where VR + IA outperformed OV + IA in all Q1-Q5 dimensions. The slight mean difference (M_diff = 1,425.69) is likely due to the somewhat better dynamic effects, subject matter, and visual complexity of VR content videos. However, the difference is not statistically or practically significant. In addition, the greater variability of OV (SD = 16,233.89) diminishes this slight significance. This finding does not support the initial hypothesis that VR environments significantly increase gaze attention to visual contents. Since VR content video is 2D, the same screen and

interaction occur in it as in regular video, without the real VR features that include 360° immersive vision, head tracking, and user interaction with spatial awareness (Bhowmik, 2024). The initially conceived VR, intended as a real immersive experience, could not be studied with respect to attention in this experiment, which showed only VR content videos. However, IA enhances visual engagement through cross-modal attention guiding features, which align with auditory-led attention research (Cheng et al., 2016). When stripped of its immersive and interactive features, VR content video did not significantly attract attention compared to regular videos on a flat screen ($d < 0.079$). The results were also influenced by data distribution, rather than the benefits of VR content video, with the OV group, which included participants interested in specific videos, having a higher average rank than the VR content video group. This reflects the nonlinear trigger of Flow theory. When participants' personality traits (high concentration), their level of understanding (high education) and the characteristics of the stimulation (music and video matching) correspond with individual and environmental thresholds, their fixed gaze time can differ from the norm and cause "time distortion" even if the interaction is RA and the environment is OV (Wang et al., 2025). Responses from the group using VR content video were more concentrated, with no value extremes, resulting in a lower ranking compared to the OV group. The combination of various visual elements from VR and gamified content, along with AI, effectively stimulated emotional resonance (Q1/Q4/Q5) and cognitive involvement with the subject being presented (Q2/Q3). This experience is needed to change gaze times and serves to support the media psychology argument that presence results from the cognitive and emotional factors involved with an immersive medium rather than solely from the behavioral indicators (Bujić et al., 2023). For example, some OV + IA participants showed high immersion (Q2 = 3.2) but had below-average gaze time, possibly because they used IA's narrative sound effects to recreate visual scenes, reflecting different mental strategies. This explains the very high gaze time in the OV group. The primary limitation of this study is the definition of VR content video. The VR video clips used were 2D video clips, which would negate all the true VR elements and would not allow for the investigation of the attentional benefits of VR. Although VR provides a greater presence than normal videos, when compared to a 2D video, it does not inherently demand attention. The impact of VR is maximized when immersive sound is included with the pictures, in line with Chion's conceptual framework, which states that sound contributes information, meaning, narrative dimensions, structural values or expressive values, thus creating added value (Chion, 2019).

Practical significance and potential applications

For VR creators facing limited budgets and resources, this study provides a clear strategy. Instead of constantly optimizing visual fidelity and complex modeling, they can focus more on optimizing immersive audio. Especially when mainstream push platforms have 2D screens, immersive audio can ensure the experience and emotional value. Moreover, this study shows that even without head-mounted VR devices (which are expensive and have high barriers), users can still obtain an immersive experience by having stereo headphones. This provides a more effective path for promoting high-quality immersive experiences on common hardware devices such as mobile phones and laptops. Finally, our findings challenge the visual-centered paradigm. We advocate for the importance of spatial audio, which is regarded as the core immersive driving factor in the early concept design stage. This "audio-first" strategy is applicable to scenarios such as education, virtual travel, and online games that require high concentration and emotional connection.

Limitations and future directions

Although this study provides a new perspective on how immersive audio in a pseudo-VR design might influence gaze interaction duration and emotional resonance, it has limitations. The "VR" stimuli were implemented on a non-immersive 2D screen, which meant there was really no depth experience, head tracking, or interactive components related to the VR itself. Second, the consistent order of the stimuli may have complicated the effect by incorporating factors such as systematic learning and fatigue. Nevertheless, we believe that the experiment is still feasible for the initial observation of phenomena under specific sequences: the combination of immersive audio and VR content videos has the potential to enhance the experience. Future research will adopt a completely balanced (Latin square design) or randomized stimulus sequence to replicate and verify the effects observed in this study. Besides, we have never made a standard check of participants' listening ability, especially their ability in bilateral hearing balance. Because of the reliance on accurate perception of immersive audio in our experiment, some participants may possess undetected auditory impairments, which could influence the accuracy of their responses to specific immersive audio stimuli in terms of emotional and physiological reactions. Therefore, future research will incorporate an audio frequency test into the participant selection process. Notwithstanding this constraint, the results presented in this work may indicate the potential for immersive audio and suggest significant improvements to the methodology. In addition, this study did not survey participants' appreciation of "immersive experience" or their expectations of the situation. These two factors may account for data that cannot be quantified. In this study, multiple items were set in the questionnaire survey to approach the core concept of "presence" from various aspects, avoiding simplification. Secondly, the experimental task itself (experiencing a specific video in similar headphones and environment) provided participants with a standard and highly immersive framework. Therefore, the scores in a particular context can still ensure the objectivity of the data. Before the experiment, we also did not measure individual flow tendencies, meaning individual differences—particularly extreme responses related to cognitive load—were not accounted for. Lastly, most participants were from China, so the cultural diversity necessary for broad generalization might be affected by differences in media perception, emotional expression, and spatial understanding.

Future direction

Future research could convert non-verbal cues, such as emotional signals, into recognizable sound symbols that trigger subconscious anxiety in viewers. It could also add more evaluation dimensions to current indicators by creating a more comprehensive cross-modal experience assessment system—such as sound-image synchronization rate and physiological arousal levels—to analyze the overall effect of multi-sensory experiences on viewers' emotional responses (Kobayashi et al., 2015). Additionally, it should include the "Flow susceptibility" scale (similar to the DFS-2) or utilize pre-experiment baseline tests to account for individual differences (Procci et al., 2012). To address the issue of defining the core idea within the limitation, we could determine the meaning of "immersion" in our research to ensure all participants have a common understanding.

Conclusion

Immersive audio is crucial for enhancing media experiences, especially in VR content video, which requires the combination of immersive audio to create a synergistic effect where 1 + 1 equals more than 2. Without VR devices' immersive technology, VR content video is similar to regular videos in capturing attention. However, its distinctive characteristics, such as an improved sense of motion, can significantly contribute to the subjective feeling of immersion when combined with immersive audio. This conclusion calls into question the traditional "visual-focused" design principles, establishing sound as the primary element in the multisensory media of the future and offering a crucial basis for multisensory content across multiple platforms. Without this sort of immersive audio, there is virtually no VR experience. The use of immersive audio (which entails a simplified feedback mechanism) and virtual-reality enhanced videos (which heighten the sense of enclosure) allows for more easily delivering the Flow State because they decrease the matching threshold. However, outliers of conventional audio and regular videos show that when individual factors (such as preference bias) or stimulus randomness (the match degree of music and video) compensate for environmental deficiencies, the threshold may still be surpassed.

Inspirations for virtual reality film audio design

During the pre-production stage, research shows that the contribution of immersion audio to presence is greater than that of visual media complexity (OV + IA > VR + RA). High-order Ambisonics or object-based audio (such as Dolby Atmos) is used to achieve hemispherical sound field coverage. In the pre-test phase, 3D scans of the user's earlobes are collected through HRTF (Head-Related Transfer Function) personalized calibration, creating a personalized head-related transfer function database that calculates the sound source's direction in real-time. Additionally, the sound source direction is calculated in real time based on the viewer's head rotation, and with heart rate detection, the rhythm of the BGM can be dynamically changed. An additional mapping model connects heart rate to various music parameters, allowing for modification based on specific heart rates during video production. This ensures the video's effects more accurately mirror the participants' psychological and physiological conditions. When combining platform Unity 2021.3, Wwise 2022.1, and Bluetooth heart rate sensors (PulseSensor), if aim to show footsteps' direction cues the character's actions and positions, and visual emotions such as tension and spatial sense are visualized, then, based on Raycast detection of ground material, corresponding events are triggered, and Wwise can automatically handle random repetitive playback. Additionally, Wwise can simulate various space sounds (such as valley reverberation or indoor reverberation) by setting attenuation curves and spatial movement, and manage multiple language tracks to switch language packages dynamically, better meeting the needs of different groups.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Ethics statement

The studies involving humans were approved by the Ethics Committee of Nanchang University. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

QH: Writing – original draft, Visualization, Writing – review and editing, Methodology, Investigation. ZL: Writing – review and editing, Methodology, Supervision, Conceptualization.

Funding

The authors declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frvir.2025.1691405/full#supplementary-material

References

Ananthabhotla, I., Ramsay, D. B., Duhart, C., and Paradiso, J. A. (2021). Cognitive audio interfaces: mediating sonic information with an understanding of how we hear. *IEEE Pervasive Comput.* 20 (2), 36–45. doi:10.1109/MPRV.2021.3052659

Bhowmik, A. K. (2024). Virtual and augmented reality: human sensory-perceptual requirements and trends for immersive spatial computing experiences. *J. Soc. Inf. Disp.* 32 (8), 605–646. doi:10.1002/jsid.2001

Bujić, M., Salminen, M., and Hamari, J. (2023). Effects of immersive media on emotion and memory: an experiment comparing article, 360-video, and virtual reality. *Int. J. Human-Computer Stud.* 179, 103118. doi:10.1016/j.ijhcs.2023.103118

Cheng, C. H., Chan, P. Y. S., Niddam, D. M., Tsai, S. Y., Hsu, S. C., and Liu, C. Y. (2016). Sensory gating, inhibition control and gamma oscillations in the human somatosensory cortex. *Sci. Rep.* 6 (1), 20437. doi:10.1038/srep20437

Chion, M. (2019). "The audiovisual scene," in *Audio-vision: sound on screen* (New York Chichester, West Sussex: Columbia University Press), 66–97. doi:10.7312/chio18588-006

Csikszentmihalyi, M. (1975). *Beyond boredom and anxiety*. Jossey-Bass.

Csikszentmihalyi, M. (2014). *Applications of flow in human development and education*. Dordrecht: Springer.

Eggermont, J. J. (2019). *The auditory brain and age-related hearing impairment*. Academic Press. Available online at: https://lccn.loc.gov/2018959311.

Engeser, S., and Rheinberg, F. (2008). Flow, performance, and moderators of challenge-skill balance. *Motivation Emot.* 32 (3), 158–172. doi:10.1007/s11031-008-9102-4

Fritz, C. O., Morris, P. E., and Richler, J. J. (2012). Effect size estimates: current use, calculations, and interpretation. *J. Exp. Psychol. Gen.* 141 (1), 2–18. doi:10.1037/a0024338

Gallese, V. (2005). "Being like me: self–other identity, mirror neurons, and empathy,". *Perspectives on imitation: from neuroscience to social science*. Editors S. Hurley and N. Chater (Cambridge, MA: MIT Press), 1, 101–118. doi:10.7551/mitpress/9780262083369.003.0005

Howlett, J. R., and Paulus, M. P. (2017). Individual differences in subjective utility and risk preferences: the influence of hedonic capacity and trait anxiety. *Front. Psychiatry* 8, 88–8. doi:10.3389/fpsyt.2017.00088

Husselman, T. A., Filho, E., Zugic, L. W., Threadgold, E., and Ball, L. J. (2024). Stimulus complexity can enhance art appreciation: phenomenological and psychophysiological evidence for the pleasure-interest model of aesthetic liking. *J. Intell.* 12 (4), 42. doi:10.3390/jintelligence12040042

Jackson, S. A., and Marsh, H. W. (1996). Development and validation of a scale to measure optimal experience: the flow state scale. *J. Sport Exerc. Psychol.* 18 (1), 17–35. doi:10.1123/jsep.18.1.17

Kobayashi, M., Ueno, K., and Ise, S. (2015). The effects of spatialized sounds on the sense of presence in auditory virtual environments: a psychological and physiological study. *Presence Teleoperators Virtual Environ.* 24 (2), 163–174. doi:10.1162/PRES_a_00226

Kujawska-Lis, E. (2000). Symbolism and imagery in "Heart of Darkness" and Apocalypse now. *Acta Neophilol.* (II), 133–153. CEEOL - Article Detail.

Lehman-Wilzig, S., and Cohen-Avigdor, N. (2004). The natural life cycle of new media evolution: inter-media struggle for survival in the internet age. *New Media and Soc.* 6 (6), 707–730. doi:10.1177/146144804042524

Li, X., and Xie, X. (2023). "Avatar: Shui zhi dao": "Ji shu qi guan" dian ying de te xiao cheng xian yu tuo yan ce lüe ["Avatar: the Way of Water" the special effects presentation and extension strategies of "technical spectacle" film]. *Adv. Motion Pict. Technol.* (2), 37–49. doi:10.3969/j.issn.1673-3215.2023.02.006

Parvizi-Wayne, D., Sandved-Smith, L., Pitliya, R. J., Limanowski, J., Tufft, M. R., and Friston, K. J. (2024). Forgetting ourselves in flow: an active inference account of flow states and how we experience ourselves within them. *Front. Psychol.* 15, 1354719–1354726. doi:10.3389/fpsyg.2024.1354719

Procci, K., Singer, A. R., Levy, K. R., and Bowers, C. (2012). Measuring the flow experience of gamers: an evaluation of the DFS-2. *Comput. Hum. Behav.* 28 (6), 2306–2312. doi:10.1016/j.chb.2012.06.039

Romero-Fresco, P., and Fryer, L. (2013). Could audio-described films benefit from audio introductions? An audience response study. *J. Vis. Impair. and Blind.* 107 (4), 287–295. doi:10.1177/0145482X1310700405

Shang, H., Liu, X., Liang, Z., Zhang, J., Hu, H., and Guo, S. (2025). United minds or isolated agents? Exploring coordination of LLMs under cognitive load theory. *arXiv Prepr. arXiv:2506.06843*, 4–9. doi:10.48550/arXiv.2506.06843

Sherry, J. L. (2004). Flow and media enjoyment. *Commun. Theory* 14 (4), 328–347. doi:10.1111/j.1468-2885.2004.tb00318.x

Somarathna, R., Bednarz, T., and Mohammadi, G. (2022). Virtual reality for emotion Elicitation–a review. *IEEE Trans. Affect. Comput.* 14 (4), 2626–2645. doi:10.1109/TAFFC.2022.3181053

Stock, K. (2009). Fantasy, imagination, and film. *Br. J. Aesthet.* 49 (4), 357–369. doi:10.1093/aesthj/ayp030

Swann, C., Keegan, R. J., Piggott, D., and Crust, L. (2012). A systematic review of the experience, occurrence, and controllability of flow states in elite sport. *Psychol. Sport Exerc.* 13 (6), 807–819. doi:10.1016/j.psychsport.2012.05.006

Ulrich, M., Keller, J., Hoenig, K., Waller, C., and Grön, G. (2014). Neural correlates of experimentally induced flow experiences. *Neuroimage* 86, 194–202. doi:10.1016/j.neuroimage.2013.08.019

Wang, D., Rhee, C., and Park, J. (2025). Exploring the role of time distortion in psychological well-being: the impact of evocative VR content. *Behav. and Inf. Technol.*, 1–20. doi:10.1080/0144929X.2025.2523444

Webster, J., Trevino, L. K., and Ryan, L. (1993). The dimensionality and correlates of flow in human-computer interactions. *Comput. Hum. Behav.* 9 (4), 411–426. doi:10.1016/0747-5632(93)90032-N

Yu, Z., and Lo, C. H. (2023). "The emotional impact of camera techniques in cinematic virtual reality: examining frame shots and angles," in *Proceedings of the future technologies conference*, 543–563. doi:10.1007/978-3-031-47454-5_38