frontiers | Frontiers in Virtual Reality

ORIGINAL RESEARCH
published: 22 June 2022
doi: 10.3389/frvir.2022.874054

Eye Movements During Comprehension in Virtual Reality: The Influence of a Change in Point of View Between Auditory and Visual Information in the Activation of a Mental Model

Véronique Drai-Zerbib[1]*, *Léa Bernigaud*[1], *Alexandre Gaston-Bellegarde*[1], *Jean-Michel Boucheix*[1] and *Thierry Baccino*[2]

[1]Laboratory LEAD-CNRS, UMR5022, Université Bourgogne, Dijon, France, [2]LUTIN, University Paris 8, Paris, France

OPEN ACCESS

Edited by:
Vsevolod Peysakhovich,
Institut Supérieur de l'Aéronautique et de l'Espace (ISAE-SUPAERO), France

Reviewed by:
Pierre Raimbaud,
Inria Rennes—Bretagne Atlantique Research Centre, France
Daniel Rene Mestre,
Aix-Marseille Université, France

***Correspondence:**
Véronique Drai-Zerbib
Veronique.Drai-Zerbib@u-bourgogne.fr

Specialty section:
This article was submitted to Technologies for VR, a section of the journal Frontiers in Virtual Reality

Received: 11 February 2022
Accepted: 30 May 2022
Published: 22 June 2022

Citation:
Drai-Zerbib V, Bernigaud L, Gaston-Bellegarde A, Boucheix J-M and Baccino T (2022) Eye Movements During Comprehension in Virtual Reality: The Influence of a Change in Point of View Between Auditory and Visual Information in the Activation of a Mental Model. Front. Virtual Real. 3:874054. doi: 10.3389/frvir.2022.874054

This paper provides new research perspectives in the field of multimodal comprehension (auditory crossing visual information) by using immersion and incorporating eye tracking in a virtual reality environment. The objective is to investigate the influence of a change in narrative perspective (point of view) during the activation of a mental model underlying comprehension between visual and auditory modalities. Twenty-eight participants, equipped with a headset SMI HMD HTC eye-tracking 250 Hz watched 16 visual scenes in virtual reality accompanied by their corresponding auditory narration. The change in perspective may occur either in the visual scenes or in listening. Mean fixations durations on typical objects of the visual scenes (Area of Interest) that were related to the perspective shift were analyzed as well as the free recall of narratives. We split each scene into three periods according to different parts of the narration (Before, Target, After), the target was where a shift in perspective could occur. Results shown that when a visual change of perspective occurred, mean fixation duration was shorter (compared to no change) for both Target and After. However, when auditory change of perspective occurred, no difference was found on Target, although during After, mean fixation duration was longer (compared to no change). In the context of 3D video visualization, it seems that auditory processing prevails over visual processing of verbal information: The visual change of perspective induces less visual processing of the Area of Interest (AOIs) included in the visual scene, but the auditory change in perspective leads to increased visual processing of the visual scene. Moreover, the analysis showed higher recall of information (verbatim and paraphrase) when an auditory change in perspective was coupled with no visual change of perspective. Thus, our results indicate a more effective integration of information when there is an inconsistency between the narration heard and viewed. A change in perspective, instead of creating comprehension and integration difficulties, seems to effectively raise the attention and induce a shorter visual inspection. These results are discussed in the context of cross-modal comprehension.

Keywords: virtual reality, eye tracking, comprehension, multimodality, shift in narrative perspective

1 INTRODUCTION

All complex real-life situations need a full and adequate comprehension of their elements to provide an adapted answer and behavior. The comprehension process is based on a sophisticated system of cognitive elaboration and needs the activation of a coherent mental model from the start to the end of an event or a situation (Kintsch, 1998). The concept of mental model refers to a body of knowledge that humans have acquired and which are used to understand simple or complex situations. The individual builds an internal model (mental model) of external reality, including the various associated possibilities, then chooses the most suitable action among several alternatives that will allow them to understand and react to situations, while using prior knowledge (Johnson-Laird, 1980; Johnson-Laird, 1993).

Classically, comprehension has been mostly studied for text reading but the process is very comparable for understanding a visual scene and one of the general model of comprehension (Construction/Integration model) for text may be applied easily for such situations. The C/I model developed by Kintsch (Kintsch, 1988; Kintsch, 1998) for text understanding is enough general to be applied to visual scenes. Briefly, the comprehension needs two processes (Construction and Integration) and three levels of representation (surface, propositional, situation model) from which the understanding may emerge. Comprehension starts as soon as reading begins, in such a way that a mental representation of text meaning is built gradually (called mental model or situational model), getting richer and richer as new information is supplied by the text. The comprehension is therefore a dynamic process that updates with new information and links new information with previous knowledge. One important property of this model is to claim that comprehension emerges only after several operations of coherence. Local coherence at word level when it comes to linking adjacent propositions of the text, and global coherence when it comes to adjust knowledge from the text base (and after application of semantics rules) to the specific knowledge of the reader stored in memory. Local and global coherence are mainly due to inference mechanisms that allow linking different parts of text even if the information is not explicitly expressed. This linking is carried out by information extracted from the encyclopedic reader's memory and from the mental model built along text reading. When the mental model is known by the reader (when analogous situations were encountered), the comprehension is achieved. The misunderstanding arises when the current mental model does not correspond to any semantic structure stored in memory or when coherence operations (inferences) failed. A large body of research has taken an interest in the processes used by readers to maintain coherence as they build an integrated mental representation of the information drawn from the text and from their own knowledge base. Many discourse features contribute to maintaining coherence, including referential (Kintsch and Van Dijk, 1978; Kintsch, 1988), and causal relations (Fletcher et al., 1990). Research has shown that a break in the coherence of a narrative brought about by the introduction of a new theme (Lorch et al., 1987; Hyönä and Niemi, 1990; Hyönä, 1994; Hyönä, 1995) or a perspective shift (Millis et al., 1995; Baccino and Pynte, 1998) increases reading time on the line in question. A longer reading time is generally considered to reflect the occurrence of an integration process. The extra processing time at topic shift or coherence break is spent on encoding that new piece of information, integrating it with the previous representation of text by building a new sub-structure. Evidence for these processes come from a variety of experiments of reading and comprehension of texts with perspective shift, topic shift, use of adverb informing readers about the narrative structure (Gernsbacher, 1990; Gernsbacher, 1995; Gernsbacher, 1996).

Initially focused in the verbal representational system, mental models have gradually adapted to the addition of images sing the pictorial representation system in multimedia documents that have gradually emerged in everyday life. Indeed, nowadays, in the age of digitalization, more than ever a mix of multimodal information (sound, text, image, and video) can be displayed on screen or in virtual environments, in full immersion. Several studies aimed to understand how the comprehension in the context of multimodal information works (Johansson et al., 2012; Drai-Zerbib and Baccino, 2017). When there are different sources of information, connections between information established by comprehension aim to create a unique, coherent, integrated mental representation (mental model) by taking into account relevant information from verbal and visuospatial external sources (Fayol and Gaonac'h, 2003). Comprehension could be improved by multimedia, when two modalities are at work: this improvement has been coined the multimedia principle (Mayer, 2014; Mayer et al., 2020). The addition of illustration strengthens comprehension, supports memorization and provides information consolidation. In other words, we could learn better when the document is enriched with an image or an animation (e.g. (Mayer, 2014). Images and videos, when correctly designed and integrated into a presentation, facilitate the comprehension and memorization: a higher number of semantic units are memorized when learners have been exposed to both oral messages and animations rather than separated auditory and visual information (Guichon and McLornan, 2008). However, a semantic congruency of auditory and visual information seems mandatory to improve the recognition of items and the comprehension process (Meyerhoff and Huff, 2016). This is also the case during visual search, where the characteristic sounds make look at the target objects more quickly when they are semantically congruent (Iordanescu et al., 2008). Therefore, the semantic matching of information across sensory channels is a crucial factor influencing cross-modal integration (Senkowski et al., 2008). More generally, the semantic matching may be done at different levels of processing, at the semantic level during text reading activity (Kintsch and Van Dijk, 1978; Kintsch and Mangalath, 2011) or (more appropriately here) at referential level (mental models in Johnson-Laird perspective, 1980) when the semantic image representation have to be aligned with the discourse representation.

The semantic coherence, through the narrative perspective has been previously studied to investigate the impact of a change in point of view on comprehension (Baccino and Pynte, 1998). In narration, the point of view establishes the relationship between the narrator and the action (description of characters and events) (Genette, 1972). The narrative perspective can be focused in two different angles: an internal, point of view, where the narrator merges with one or more characters. The story is told through a subjective vision of time and space, with the use of the first person singular (I). The viewing position is inside the story. An external point of view, where the narrator is invisible. The story is told in an objective manner, with the use of the third personal singular (he/she/they). The viewing position is outside the story. Previous research (Gu and Tse, 2016) has shown that comprehension is most enhanced when the narrative perspective is focused on the internal point of view during reading. Moreover, to retrieve autobiographical events, the first person singular (I) leads to follow a narrative perspective in which the person is an actor, while the third personal singular (he/she/they) refers to a narrative perspective in which the person is the observer (Gu and Tse, 2016). Indeed, the point of view taken by the narration contributes to the narrative coherence. The point of view is an inherent feature of a narration (Black et al., 1979) and is a way of regulating information, which refers to narrative instances in which the individual is inscribed. To summarize, the narrative perspective is associated with a point of view (first or third person) taken during the narration, and affects the comprehension, memorization and coherence because this narrative perspective influences the relationship between the subject and the events.

Although most narrations adopt the same narrative perspective from the start to the end of the story, an author may change the perspective, which brings the change of the point of view. When the narrator and the character remain distinct from the start to the end of the narration there is no change in perspective. Therefore, sentences are read more quickly and more accurately remembered because the main character's perspective is adopted and maintained throughout (Black et al., 1979; Millis et al., 1995; Graesser et al., 1997). On the contrary the change in narrative perspective during the story influences memory capacities and remembering (Gu and Tse, 2016). In reading, a shift in narrative perspective, when the narrator becomes the character during the narration (a change from an external point of view to an internal point of view) disrupts the comprehension and integration process (Black et al., 1979). In this case, a referential difficulty is encountered. Therefore, an additional cognitive effort is necessary to integrate the new information into the current memory representation, in order to update adequately the mental model (Albrecht and O'Brien, 1993; Kintsch and Van Dijk, 1978). Consequently a longer reading time in the area in which the change of perspective took place was observed (Baccino and Pynte, 1998). A recent study using concurrent recordings of eye movements and head motion showed fluctuation in cognitive engagement when a change in narrative perspective occurs during reading. Longer fixation durations, longer reading time and a higher stability of posture with decreasing head motions when perspective changed reflected this extra cognitive engagement (Ballenghein and Baccino, 2019). Logically, even if a part of the reading process is automated, the continuous elaboration of a situation model along the reading (as any other cognitive activity) requires the allocation and maintenance of attentional resources (Kahneman, 1973). Thus, when more attentional resources are allocated to the information processing (Cowan, 1995), it is reflected by longer fixation duration. This is the case when new information requires readers to spend additional resources to update adequately the activated mental model (Hyönä and Lorch, 1998).

As other media than books can be used for narration, in this paper we were interested to investigate the shift of perspective, when the media request auditory and visual modality integration. Recent research on multimedia comprehension used immersive virtual reality—VR (Makransky et al., 2019a; Makransky et al., 2019b; Makransky et al., 2020; Makransky and Petersen, 2021; Petersen et al., 2022). In many of these studies, two presentation displays are compared: a full immersive condition (VR, with a Head Mounted Device -HMD-) is contrasted with an on-screen PC video condition presenting the same images in 2D. The results showed, as stated by Makransky et al., 2019a, that *"Adding immersive virtual reality . . ..causes more presence but less learning"*. Learning and comprehension performances were often found higher in the PC video condition than in the VR condition. Further, in several experiments EEG investigations during learning in both conditions showed a higher cognitive workload in the VR condition than in the PC-Video condition (Makransky et al., 2019b; Baceviciute et al., 2020; Baceviciute et al., 2021). Such results may suggest that, 3D with visual richness in VR, increase perceptual, cognitive load and may be distraction during comprehension of verbal information compared to non-immersive on-screen presentations.

However, recent development of virtual reality (VR) allows conducting experiments under more ecologic conditions to understand cognitive processes (Jungnickel and Gramann, 2016) and to reach high immersion of the subject with the virtual environment (Clay et al., 2019). Further, eye tracking in VR is a relatively new and a promising technique. The combination of eye tracking and VR makes it possible to calculate the gaze of the subject in 3D space and observe where the subject is looking at during the session (Clay et al., 2019). Thus, we investigated, with eye tracking embedded in VR, the change in narrative perspective in a multimodal virtual environment, crossing auditory and visual representation of the narration. Our aim was to observe the influence of the point of view in the activation of a mental model during cross modal narration comprehension. We used a virtual reality HMD device with integrated eye tracking to observe eye movements during simultaneously listening and exploring the 3D visual scene corresponding to the narration with or without change in perspective. Eye movements' data should provide information about the activation of a mental model during comprehension and the update of this mental model between visual and auditory modalities into the immersive environment. Furthermore, using this technique, the aim of our study was to bring new research perspectives in the field of comprehension in dynamic situation, using virtual reality environment to provide high immersion in the scene.

What will be the impact of a shift in narrative perspective between the auditory and visual information on the activation of a mental model and comprehension? How will be processed the change in narrative perspective (from an external point of view to an internal point of view) between the auditory narration and the corresponding visual scene displayed in virtual reality?

Based on previous studies in reading and change in perspective, we should assume that comprehension would be disrupted in cases where a shift in perspective occurred. Participants will need to assess the consistency of the change with respect to what precedes and modify the mental representation of the narration accordingly. The lack of coherence between the simultaneous virtual environment displayed and the listened narration should induce longer mean fixation durations (H1). In particular mean fixation durations should be longer in the target area of interest (AOI), when there is a mismatch between the perspective of the auditory narration and the perspective of the visual stimuli (H2). The free recall is also expected to have poorer content (less verbatim recalled) (H3). In other words, a change in perspective might trigger comprehension and integration difficulties because of inconsistency between the simultaneously narration heard and VR scenes visualized.

2 MATERIALS AND METHODS

2.1 Participants

Forty-one volunteers were recruited at University Bourgogne Franche-Comté. They gave their written consent and were compensated by credit course. Eleven participants were excluded from the eye-tracking data set due to weak eye-tracking ratio (<40%). The final data set consisted in 28 participants (22 female), mean age 21 years. They were all native French speakers, right-handed and reported normal or correct-to normal vision. Their task was to visualize 16 scenes in virtual reality and summarize orally the narration. They were naïve to the purpose of the experiment. A power analysis (G*Power 3 (Erdfelder et al., 1996; Faul et al., 2009)) revealed a total sample size of 18 participants would be required to obtain statistical power with (d = 0.25) and power (1−β) set at 0.95 (α = 0.05) using a F test (Repeated measures ANOVA, within factors).

2.2 Materials

2.2.1 Narrations Used for Creating VR Stimuli

Narrations were adapted from text material (Baccino and Pynte, 1998), also used in Ballenghein and Baccino (2019). It consisted of sixteen selected stories containing two paragraphs written in French. Twelve texts were used to cross the experimental conditions, with six texts including a shift in perspective (3rd person/1st person) between the first and the second paragraph, six texts maintaining the same perspective (3rd person). Two texts were used as filler texts (never shifting), and two were training text, allowing the participant to get used to the experimental design (familiarization to VR audio-visual presentation). All texts were equivalent in number of words, lines and paragraphs (44 words on average, seven lines long, two paragraphs). The first paragraph defined the general setting by giving a very vague description of the place or time of the main action, which was introduced at the beginning of the second paragraph. The last sentence of the first paragraph mentioned the character who would be the main actor in the narrative. The first sentence of the second paragraph is the target sentence where change in perspective might occur. An example of those texts (see below) shows the **target sentence** (in bold) containing either a change in perspective or no change in perspective between the first and second paragraph (translation in brackets).

1st Paragraph

Le soleil luit vaguement sur les rochers. C'est une heure étrange, où les bruits ne se perçoivent que comme amortis par un voile opaque, et les gens ne semblent pas avoir de contour ni de réalité. Un personnage se détache cependant du groupe.

(The Sun shone vaguely on the rocks. It was a strange hour, one where noises seemed to be muffled by an opaque screen, and even the people seemed to have neither shape nor reality. One person, however, stood out from the group.)

2nd Paragraph

Perspective-shift condition ((je - I)

Tristement, je débarque du navire. Au milieu de gens pressés, **je** descends sur le quai avec fièvre. **je** n'aime pas les ports, et pas davantage cette ville perchée sur la mer du Nord.

(**Sadly, I got off the ship**. In the middle of the hurrying crowd, **I** feverishly went down to the docks. **I** did not like ports, no more than this town perched over the North Sea).

No Change (il–he) in perspective condition

Tristement, il débarque du navire. Au milieu de gens pressés, **il** descend sur le quai avec fièvre. **Il** n'aime pas les ports, et pas davantage cette ville perchée sur la mer du Nord.

(**Sadly, he got off the ship**. In the middle of the hurrying crowd, **he** feverishly went down to the docks. **He** did not like ports, no more than this town perched over the North Sea).

From this text material, sixteen video files for VR display were created with typical objects of the narration included in the visual scenes. Twelve video files were used to cross the six experimental conditions of audio and visual change (shift) in perspective as detailed **Table 1**. The audio versions of the narrations were recorded by the experimenter (Audacity™ software) and combined with VR scenes created with Unity 2018.5f1 (see section *virtual environment*). The experimental design was a within-subject design with two factors: audio perspective shift with two levels: Shift/no Shift; visual perspective shift, with three levels: Shift/no Shift/Grey mask. The Grey mask condition is a control condition. At the end of each video visualization, participants had to orally summary the narration. The total duration of the experimental session was about 45–60 min.

2.2.2 Eye-Tracking and VR Setup

A complete typical experimental setup of eye tracking in virtual reality (VR) was used for this experiment. The *VR-Hardware set*

TABLE 1 | Presentation of the six experimental conditions crossing.

Experimental conditions		Visual perspective		
		Shift (He/I)	No Shift (He/He)	Grey mask (Control)
Audio perspective	Shift (He/I)	X	X	X
	No Shift (He/He)	X	X	X

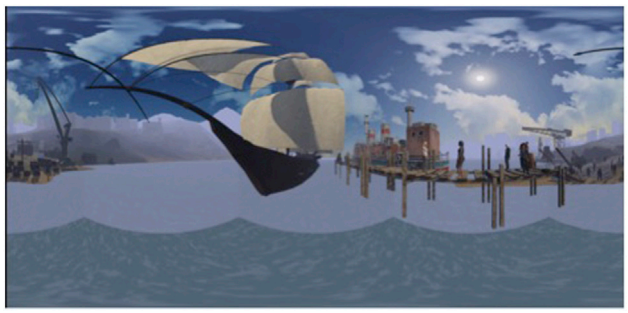

FIGURE 1 | Start of the 360°VR scene related to the narration cited in example.

used the SMI tethered eye-tracking virtual reality (VR) head-mounted dispay (HMD), powered by HTC and Vive, with an integrated binocular eye tracker from SMI. The SMI Eye Tracking HMD is thus a HTC Vive with an embedded SMI eye-tracking device. The HTC Vive has dual Active-Matrix Organic Light-Emitting Diode (AMOLED) screens, each with a 3.6-inch diagonal with a resolution of 2,160 × 1,200 pixels (1,080 × 1,200 pixels per eye). The SMI embedded eye-tracking device tracks both eyes simultaneously with a sampling rate of 250 Hz and with a spatial accuracy of 0.2° (for more information see SMI HMD manual). This setup includes also two motion trackers. In this experimental setup, no interaction with the virtual world was required, thus no VR controllers were needed. The *VR-Software set* used the SteamVR plugin for viewing in the HTC Vive virtual reality headset. The area calibration was 2,160 × 1,200. The dimensions of the physical stimuli were 2,160 mm (horizontally) and 1,200 mm (vertically). Event detection was at estimated speed. The speed threshold was 40° per second, the minimum fixation duration was 50 ms, and the minimum high speed detection ratio was 22 ms. The design of the experiment as well as the calibration phase were programmed on SMI Experiment Center software 3.7. The analysis and the output of the data were carried out on the SMI BeGaze HMD HTC software 3.7. The computer running this experiment was an XMG PC computer, composed of an Intel Core i7-7700HQ @ 2.80 GHz processor and a Nvidia GeForce GTX 1070 graphics card.

2.2.3 Virtual Environment

We used Unity 2018.5f1 to create visual stimulus (scenes) for the ET-HMD, according the visual experimental conditions (shift/no shift in perspective/grey mask). We used SteamVR plugin for visualization in ET-HMD headset. The 3D models of the objects were taken from the unity asset store, the 3D Warehouse in Tromble and the website www.cadnav.com. The 3D models of the characters as well as their animations were taken from Adobe Fuse CC (beta version). We used the Unity Recorder 1.0 plugin (Unity Asset store) to capture and save data from the Unity Editor during Play mode as 360° videos within the 3D environment. Videos were captured within the environment at a constant rate of 30 frames per second in MP4 format, captured at 360° at 4,096 × 2048 resolution with stereo sound capture (**Figure 1**). The audio narrations were created with the Audacity software in. avi files according the audio experimental conditions (shift/no shift in perspective). Then video stimuli were associated with audio files in Experiment Center 3.7 software, used to run the experiment. To be able to counterbalance the order presentation of all the audio-visual stimuli, according the experimental conditions (**Table 1**), we created six experimental lists in Experiment Center. Then the order of presentation of the stimuli within each list, was automatically randomized by Experiment Center 3.7 software.

2.3 Procedure

After arriving at the laboratory, participants signed an informed consent form and completed a demographic questionnaire. Each participant was assigned to one of the six experimental lists. Then, they were comfortably installed on a chair and were equipped with the virtual reality headset (SMI Eye Tracking HMD). Eye tracking calibration consisted of tracking five randomly displayed points on the VR screen. After having received a written instruction, participants were trained on two VR audio-visual narrations (scenes/videos) at the end of which they received the fourteen randomized VR audio-visual narrations (scenes/videos). Two of them were filler scenes/videos (never shifting), twelve were experimental scenes/videos, crossing two audio condition (Shift/No Shift) and three visual conditions (Shift/No Shift/Grey mask) with two VR scenes/videos per condition (each of the visual condition, including the grey mask one, was tested in combination with all different audios). For the twelve experimental VR scenes/videos, the participants were systematically first exposed to an external point of view, during the first part of the narration. Then a shift of no shift in perspective occurred. Each VR audio-visual narration (scene/video) lasted 60 s and after each trial, the participant received the message "Tell me the story" on the VR headset. Finally, the participant had to freely recall orally for 120 s as much information as possible. Then, a new visual scene was presented (**Figure 2** illustrates the procedure).

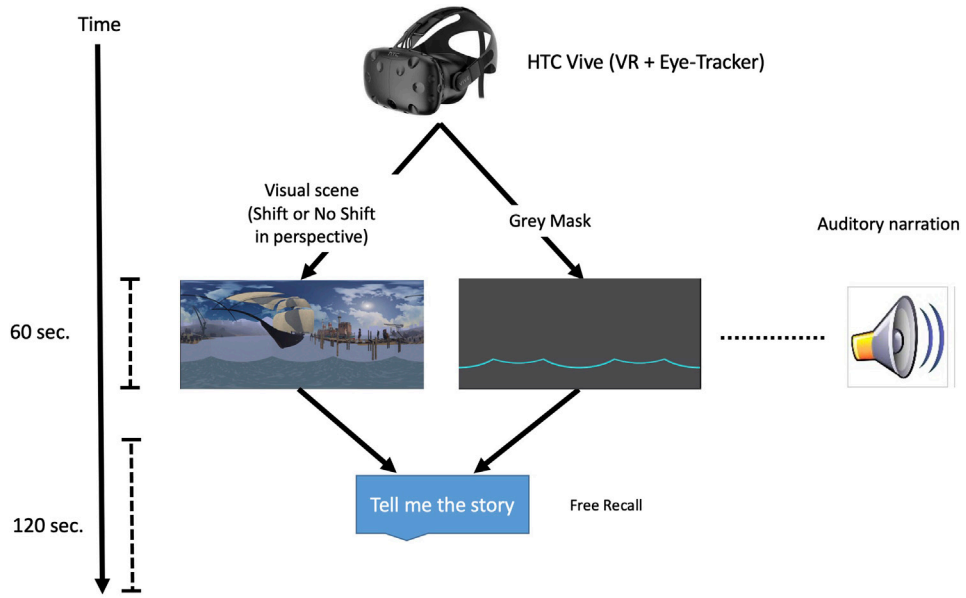

FIGURE 2 | Procedure: The participants received twelve experimental VR scenes/videos, crossing two audio condition (Shift/No Shift) and three visual conditions (Shift/No Shift/Grey mask). Each VR audio-visual narration (scene/video) lasted 60 s. After each trial, the participant received the message "Tell me the story" on the VR headset, and freely recalled orally for 120 s as much information as possible.

3 RESULTS

3.1 Data Preparation

Fixations were detected using a velocity-based algorithm (high-speed detection min ratio = 22 ms, peak velocity threshold = 40° per second, min fixation duration = 50 ms). Three AOIs related to the objects of interest constituting each scene/video were drawn. These AOIs corresponded to the typical objects of the visual scenes that were related to the perspective shift (for example, in the narration cited above, the ship, the crowd and a box on the dock were selected). Automatic Move&Morph™ function (provided in Begaze) for dynamic stimuli in video clips ensured that AOI was on the stimuli targeted across all the duration of the video (corresponding to the typical objects of the visual scenes). However, we controlled frame by frame these dynamic AOIs and it was sometimes necessary to adjust them to the experimental objects of interest. The area of the AOI is expressed in square pixels. Moreover, we split each scene into three periods, according to different parts of the narration (Before, Target, and After) for eye movements' analysis (during the data extraction in Begaze 3.7) as follows:

- Before: period that corresponded to the beginning of each scene. This was the scene/narration with an external point of view and which corresponded to the 1st paragraph (see material section above). Time = 25–30 s.
- Target: period that matched the target sentence at the beginning of the 2nd paragraph and contained either a change in perspective (he/I) or no change (he/he) between the first and second paragraph. Time = 3–5 s (It is the moment where may first intervene the disruption in case of change in perspective)
- After: period that corresponded to the rest of the text coming after the target sentence. Time = 25–30 s.

TABLE 2 | Means and standard deviations for fixation duration (ms) per area of 100 pixels as function of the period and auditory/visual perspective shift (no Shift (He/He), Shift (He/I) (N = 28).

Moment	Auditory	Visual	Mean	SD
Before	No Shift	Shift	0,1719	0,0175
Before	No Shift	No Shift	0,2227	0,0453
Before	Shift	Shift	0,1217	0,0114
Before	Shift	No Shift	0,1532	0,0205
Target	No Shift	Shift	0,1506	0,026
Target	No Shift	No Shift	0,3495	0,0888
Target	Shift	Shift	0,1317	0,0186
Target	Shift	No Shift	0,2665	0,0474
After	No Shift	Shift	0,1436	0,0188
After	No Shift	No Shift	0,4209	0,1331
After	Shift	Shift	0,13	0,0142
After	Shift	No Shift	0,1778	0,1225

3.2 Statistical Analysis

3.2.1 Eye Tracking analysis

Repeated measure ANOVA analysis (with Greenhouse-Geisser correction) was calculated on fixation durations in milliseconds (ms) per 100 pixels area (more representative of fixation area), to account for the different surfaces of the AOI (Rayner, 1998). The analysis follows a within-subject design with three factors (see

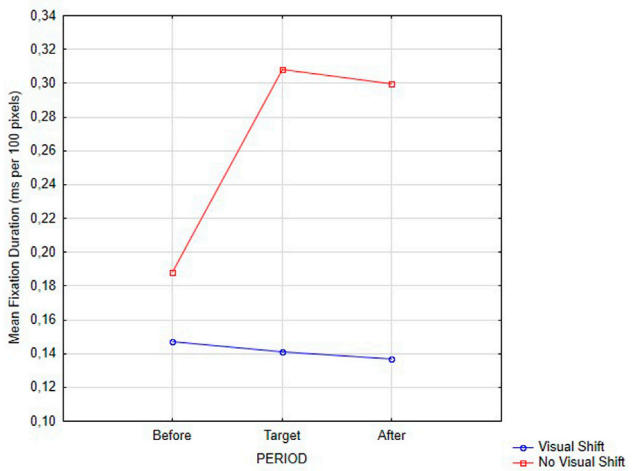

FIGURE 3 | Significant interaction between period (Before, Target, After) and VISUAL perspective shift (Visual Shift, No Visual Shift).

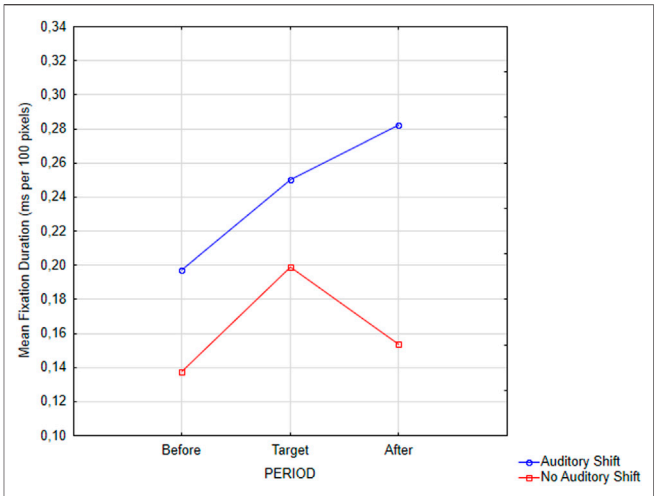

FIGURE 4 | Non-Significant interaction between period (Before, Target, After) and AUDITORY perspective shift (Visual Shift, No Visual Shift).

Table 1): audio perspective shift (Shift/No Shift) X visual perspective shift (Shift/No Shift) X periods (Before/Target/After). Analysis was not performed for the visual control condition (grey mask) because there were no stimuli to visualize. **Table 2** presents means and standard deviations results.

There was a main significant effect of the auditory perspective shift, F (1, 27) = 4.36, p = 0.046, and ηp2 = 0.12. Mean fixation duration was shorter when a perspective shift occurred compared to no shift condition. There was also a main effect of the visual perspective shift, F (1,27) = 7.019, p = 0.01331, and ηp2 = 0.21. Mean fixation duration was shorter when a perspective shift occurred compared to no shift condition. The main effect of the period was marginally significant, F (2, 54) = 2.46, p = 0.095, and ηp2 = 0.08. However, there was a significant interaction between the period and the visual perspective shift, F (2, 54) = 3,20, p = 0.049, and ηp2 = 0.11 (see **Figure 3**). Planned comparisons showed shorter fixation duration when the visual perspective shift occurred on the Target, F (1, 27) = 8,48, and p = 0.007 and After, F (1, 27) = 4,64, p = 0.04 (compared to No Shift). The comparison was logically no significant (p = 0,12) Before (as there is no video/scene difference between change and no change in perspective conditions). The interaction between the auditory change in perspective and the moment (see **Figure 4**) was not significant (F = 1).

In order to solve the time issues that are different between the Target (3 s) and After (31 s) periods, we run an ANOVA analysis separately for Target and After (we did not expect differences on Before since the perspective shift never occurs on this period and the previous analysis showed no significant effect on Before). The ANOVA was calculated on fixation durations in milliseconds (ms) per 100 pixels area, to account for different AOI surfaces. The analysis was a within-subject design with audio perspective shift (shift/no shift) X visual perspective shift (shift/no shift).

Mean Fixation Duration on the **Target** Period: There was a main effect of the visual perspective shift, F (1, 27) = 8,48, p = 0.007, ηp2 = 0.24, with shorter fixation duration when a perspective shift occurred compared to no shift. No effect of auditory perspective shift was observed (F > 1).

Mean Fixation Duration on the **After** Period: There was a main significant effect of the visual perspective shift, F (1, 27) = 4,64, p = 0.04, and ηp2 = 0.15. Mean fixation duration was significantly shorter when a perspective shift occurred compared to no shift. There was also a main significant effect of the auditory perspective shift, F (1.27) = 4.09, p = 0.053, and ηp2 = 0.13. Mean fixation duration was longer when a perspective shift occurred compared to no shift. Finally, there was a marginally significant interaction between the audio and the visual perspective shift, F (1, 27) = 3,34, p = 0.079, and ηp2 = 0.11. When there was no auditory perspective shift, fixation duration where shorter for visual perspective shift compared to no shift, F (1,27) = 4,22, p = 0.050. When there was an auditory perspective shift, there was no difference between the visual perspective conditions (F > 1).

3.2.2 Free Recall Analysis

We first analyzed the registered free recall. Some elements of the sentences reported by the participants were manually coded into five categories:

- Verbatim: number of elements containing information of the narration and reported with the words
- Paraphrases: number of elements containing information of the narration reported differently, using a synonym for a word
- Inferences: number of elements containing new information derived from the narration and from the subject's own knowledge
- Errors: number of elements containing false information, unrelated to the narration
- Visual elements: number of the reported scene visual elements

TABLE 3 | Means and standard deviations for each free recall category (N = 28).

Free recall category	Mean	SD
Verbatim	5,19	1,4
Paraphrases	3,24	0,71
Inferences	0,23	0,2
Visual elements	1,11	0,95
Errors	0,17	0,13

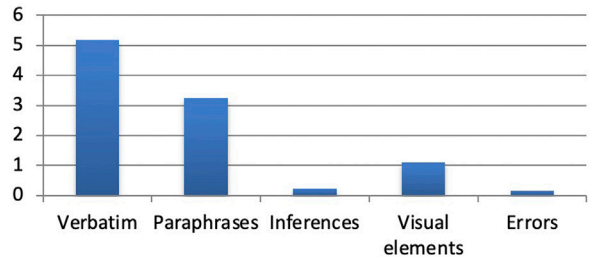

FIGURE 5 | Mean number of free recall per category.

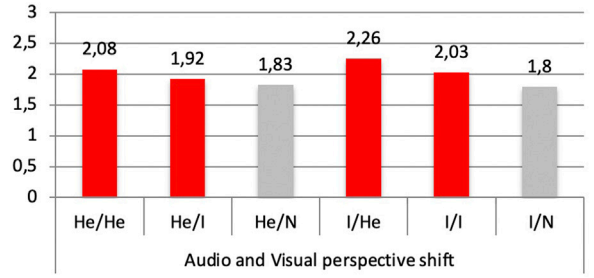

FIGURE 6 | Mean number of free recall per category according the Audio and Visual perspective shift (e.g., He/I = no audio shift in perspective combined with visual shift in perspective).

Then a repeated measure ANOVA analysis (with a Greenhouse-Geisser correction) was applied on the number of free recall. The analysis design was a within-subject design with two factors: type of recall (verbatim, paraphrase, inference, other visual object, and errors) and the six experimental conditions manipulating the perspective shift between auditory (He= No Shift; I = shift) and visual (He= No Shift; I = shift; N=Neutral) combined together (He/He, He/I, He/N, I/He, I/I, I/N). We included in this analysis the neutral (N) visual condition where a grey mask was presented in VR. Indeed, the combination with the grey mask condition were considered here since audio stimuli were present in any combinations (in opposite with previous ANOVA that must excluded this condition because of the absence of visual stimuli to track).

Overall, the number of free recalls was different across the categories, F (4,108) = 254,90, $p < 0.001$ (see **Table 3** and **Figure 5**). The number of verbatim was higher than paraphrases, visual elements, inferences or errors. The number of free recalls was also different according the six perspective shift conditions (He/He, He/I, He/N, I/He, I/I, I/N). F (5,135) = 4,68, $p < 0.001$ (**Figure 6**).

A repeated measure ANOVA analysis (with a Greenhouse-Geisser correction) was calculated to see the impact of the four perspective shift both visual and auditory (He/He, He/I, I/He, I/I) on the two main recall types: verbatim and paraphrases. As previously, we observed a significant difference between these two recall types, F (1,27) = 73.07, $p < 0.001$. Participants recalled more verbatim (M = 5.11, SD = 0.552) than paraphrases (M = 3.34, SD = 0.26). There was also a main effect of perspective shift, F (3,81) = 3.41, $p < 0.025$. Planned comparisons showed that the mean number of recall was higher in the I/HE condition (M = 4.69, SD = 0.377) compared to He/I (M = 3.85, SD = 0.375), F (1,27) = 15,04, $p < 0.001$ and also compared to I/I condition (M = 4.41, SD = 0.335), F (1,27) = 4,35, $p = 0.05$. However, there was no significant interaction between the change in perspective (visual and auditory) and the type of recall.

4 DISCUSSION

The aim of our study was to provide new research perspectives in the field of comprehension of audio-visual dynamic situations, using an immersive virtual reality environment and the technique of eye tracking. We studied how a perspective shift between auditory and visual processing could influence the information uptake during the understanding of an intermodal narration, in virtual immersion. Based on previous studies of reading and perspective shift (Baccino and Pynte, 1998), we hypothesized that comprehension would be disrupted when a perspective shift occurred between auditory and visual modalities. This disruption was supposed to induce longer fixation duration, in particular in the target moment as it is the moment of disruption, and lower performance on free recall.

In general, the change in perspective did not create understanding and integration difficulties during the recall (H3). Regardless of the presentation conditions, they managed to build a mental model and showed a high level of understanding of the narratives, as they were able to recall more verbatim and paraphrase than errors. In addition, free recall analysis also showed higher recall performance (verbatim and paraphrase) when an auditory change in perspective was presented without a visual change in perspective (I/He condition). Two different explanations are possible: 1) they were focused on visual information and disregarded auditory perspective shift, this hypothesis is consistent with the recent results by Makransky et al., 2019; Makransky et al., 2020; 2) the change in auditory narrative perspective improved their attention, required more cognitive processing and induced better memorization.

In both cases, the semantic congruency of auditory and visual information (Meyerhoff and Huff, 2016) seems less mandatory to improve the comprehension. Probably because in our study the basic story narrated remains the same even when there is a change of perspective.

Despite the lack of consistency between the simultaneously displayed virtual environment and the listened-to narration, eye-tracking analysis revealed a shorter visual inspection over the area

of interest (H1) in particular where a change in perspective occurred (H2). However, the change of perspective induces a different behavior between the auditory and visual conditions, and the period when the change of perspective occurs. When a visual change in perspective occurred, the mean duration of fixation was shorter (compared to no change) at the target and after the target. This effect on the target was not found when the perspective change is auditory. It seems that auditory processing requiring the understanding of verbal information prevails over visual processing, related to 3D video visualization. Interestingly, this effect is different from the study by Baceviciute et al., 2020 who observed that participant's learners in a VR explanatory lesson in the medical area, favored the complex visual information provided over the corresponding audio verbal information, which seemed to be more neglected than when the same verbal information was presented visually as text in the same complex virtual environment.

Further, the visual change in perspective induced less visual processing of the AOIs included in the visual scene, but the auditory change in perspective leads to rise the visual processing of the visual scene. This very relevant result may suggest that participants, in their effort to integrate the verbal and visual information tended to semantically align verbal and visual information in a very coherent multimodal information processing behavior.

In reading, longer fixation duration were observed for sentences requiring additional cognitive effort to process and integrate in a mental model (Hyönä, 1995). Previous studies investigating change in perspective for text comprehension observed that subjects read and recalled sentences more quickly and accurately when the same point of view was maintained (Black et al., 1979; Baccino and Pynte, 1998; Ballenghein and Baccino, 2019). Our results were quite different. There was more effective integration of information when a shift in perspective occurred between the narrations being listened to and viewed. A change of perspective, instead of creating difficulties of comprehension and integration, probably mobilized attention more effectively and induced a shorter visual inspection. This result suggests again a powerful text-picture integration processing which could be enhanced by the visual sense of presence in the immersive environment. Such immersive environment better does correspond to real life situations involving the continuous integration over time of verbal and visual information. Effective comprehension is also an adaptive behavior.

Clearly, the change in perspective is processed in a different manner between reading or visioning a VR scene. Verbal processing during text reading is more demanding than verbal processing during listening a narration and thus more sensible to a change of perspective. During comprehension, the mental model requires coherence to be built. In an incongruous situation, i.e., when there is a change in point of view between auditory information (narration) and visual information (VR scene) coherence seems more difficult to build (longer fixations, poorer recall) in reference to the Johnson-Laird mental model theory (1980). However, we observed the opposite. Integration and memorization seemed to have been more robust when there was a change of point of view because coherence operations required deeper processing. Of course, the semantic matching of information across sensory channels are important for cross-modal integration (Senkowski et al., 2008). However, this semantic matching may be different between a reader compared to a person visioning a story in RV immersion (different levels of processing). Reading comprehension includes three levels of coherence (surface, propositional, and referential) in order to form a mental representation (from text base and situation model—(Kintsch and Van Dijk, 1978). Thus the semantic matching may be done at a semantic level of processing during text reading activity (Kintsch and Van Dijk, 1978; Kintsch and Mangalath, 2011) and at referential level (mental models in Johnson-Laird perspective, 1980) in case of audio-visual visualization, when the semantic image representation has to be aligned with the audio representation of the narration.

Another explanation might be that when there is a change in perspective the subjects relied mainly on the auditory narration. Previous studies proposed that audio/visual integration might provide an efficient exploration of the visual scene by increasing the saliency of the corresponding objects in the scene (Van der Burg, Olivers, Bronkhorst, and Theeuwes, 2008). Meyerhoff and Huff (2016) recently showed that audio/visual scenes elicit more accurate memory performance for congruent visual and auditory information. However, violations of audio-visual synchrony do not influence memory performance.

5 CONCLUSION AND FUTURE RESEARCH

Our results provide first evidence that change in perspective processing relies on different processes for audio-visual integration in VR than reading. More studies are required and future research will allow to better understanding the mental processing of a change in point of view according the modality and the task at work. In particular, in the present study, there was no interaction with the RV stimuli, further studies should investigate whether interaction with RV might influence the change in perspective processing. Moreover, the present study observed the effect of shifting from the third to the first point of view, or keeping the third one. Further studies should investigate the cases of stable or changes from the first point of view to the third one. We could observe whether the pattern of results is similar when the narration starts with an internal point of view and shifts to the external one.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

VD-Z, TB, LB, and J-MB contributed to the conception and design of the study. VD-Z and AG-B contributed to the construction of the RV stimuli, VD-Z and LB acquired the data. VD-Z, TB, and LB performed the statistical analysis. VD-Z wrote the first draft of the manuscript. All authors contributed to the manuscript revision, and read and approved the submitted version.

FUNDING

This work was partly supported by the Region Bourgogne Franche-Comte. ANER ReQuiem Grant G039.

REFERENCES

Albrecht, J. E., and O'Brien, E. J. (1993). Updating a Mental Model: Maintaining Both Local and Global Coherence. *J. Exp. Psychol. Learn. Mem. Cognition* 19 (5), 1061–1070. doi:10.1037/0278-7393.19.5.1061

Baccino, T., and Pynte, J. (1998). Spatial Encoding and Referential Processing during Reading. *Eur. Psychol.* 3 (1), 51–61. doi:10.1027//1016-9040.3.1.51

Baceviciute, S., Mottelson, A., Terkildsen, T., and Makransky, G. (2020). "Investigating Representation of Text and Audio in Educational VR Using Learning Outcomes and EEG," in CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. https://search.ebscohost.com/login.aspx?direct=true&db=inh&AN=19936722&site=eds-live

Baceviciute, S., Terkildsen, T., and Makransky, G. (2021). Remediating Learning from Non-immersive to Immersive Media: Using EEG to Investigate the Effects of Environmental Embeddedness on Reading in Virtual Reality. *Comput. Educ.* 164, 104122. doi:10.1016/j.compedu.2020.104122

Ballenghein, U., and Baccino, T. (2019). Referential Processing during Reading: Concurrent Recordings of Eye Movements and Head Motion. *Cogn. Process* 20 (3), 371–384. doi:10.1007/s10339-018-0894-1

Black, J. B., Turner, T. J., and Bower, G. H. (1979). Point of View in Narrative Comprehension, Memory, and Production. *J. Verbal Learn. Verbal Behav.* 18, 187–198. doi:10.1016/s0022-5371(79)90118-x

Clay, V., König, P., and König, S. (2019). Eye Tracking in Virtual Reality. *J. Eye Mov. Res.* 12 (1). doi:10.16910/jemr.12.1.3

Cowan, N. (1995). *Attention and Memory: An Integrated Framework*. New York: Oxford University Press.

Drai-Zerbib, V., and Baccino, T. (2017). Effets On-Line d'un environnement musical dans la lecture de texte : analyse oculométrique. *Psychol. Française* 62 (3), 233–247. doi:10.1016/j.psfr.2014.12.002

Erdfelder, E., Faul, F., and Buchner, A. (1996). GPOWER: A General Power Analysis Program. *Behav. Res. Methods, Instrum. Comput.* 28 (1), 1–11. doi:10.3758/bf03203630

Faul, F., Erdfelder, E., Buchner, A., and Lang, A.-G. (2009). Statistical Power Analyses Using G*Power 3.1: Tests for Correlation and Regression Analyses. *Behav. Res. Methods* 41 (4), 1149–1160. doi:10.3758/brm.41.4.1149

Fayol, M., and Gaonac'h, H. D. (2003). "Aider les élèves à comprendre, du texte au multimédia," in *Editions Hachette Education*.

Fletcher, C. R., Hummel, J. E., and Marsolek, C. J. (1990). Causality and the Allocation of Attention During Comprehension. *J. Exp. Psychol.: Learn. Mem. Cogn.* 16 (2), 233–240. http://pascal-francis.inist.fr/vibad/index.php?action=search&terms=6823951

Genette, G. (1972). *Discours du Récit: Mode*. In seuil (Ed.), FIGURES III Vol. III, 183–197.

Gernsbacher, M. A. (1990). *Language Comprehension as Structure Building*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Gernsbacher, M. A. (1995). The Mechanisms of Suppression and Enhancement in Comprehension. *Can. Psychol.* 36 (1), 49–50. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=fulltext&D=ovft&CSC=Y&NEWS=N&SEARCH=00011346-199502000-00004.an

Gernsbacher, M. A. (1996). "The Structure-Building Framework: What it is, what it Might Also be, and Why," in *Models of Understanding Text*. Editors B. K. Britton and A. C. Graesser (Mahwah, NJ: Lawrence Erlbaum Associates, Inc viii, , 366), 289–311.

Graesser, A. C., Millis, K. K., and Zwaan, R. A. (1997). Discourse Comprehension. *Annu. Rev. Psychol.* 48, 163–189. doi:10.1146/annurev.psych.48.1.163

Gu, X., and Tse, C.-S. (2016). Narrative Perspective Shift at Retrieval: The Psychological-Distance-Mediated-Effect on Emotional Intensity of Positive and Negative Autobiographical Memory. *Conscious. cognition* 45, 159–173. doi:10.1016/j.concog.2016.09.001

Guichon, N., and McLornan, S. (2008). The Effects of Multimodality on L2 Learners: Implications for CALL Resource Design. Available at: https://hal.archives-ouvertes.fr/hal-00356243.

Hyönä, J. (1994). Processing of Topic Shifts by Adults and Children. *Read. Res. Q.* 29 (1), 76–90. doi:10.2307/747739

Hyönä, J. (1995). An Eye Movement Analysis of Topic-Shift Effect during Repeated Reading. *J. Exp. Psychol. Learn. Mem. Cognition* 21 (5), 1365–1373. doi:10.1037/0278-7393.21.5.a

Hyönä, J., and Lorch, R. F. (1998). Strategies for Processing an Expository Text for Recall: Evidence from Eye Fixation Patterns.

Hyönä, J., and Niemi, P. (1990). Eye Movements During Repeated Reading of a Text. *Acta Psychol.* 73, 259–280.

Iordanescu, L., Guzman-Martinez, E., Grabowecky, M., and Suzuki, S. (2008). Characteristic Sounds Facilitate Visual Search. *Psychonomic Bull. Rev.* 15, 548–554. doi:10.3758/pbr.15.3.548

Johansson, R., Holmqvist, K., Mossberg, F., and Lindgren, M. (2012). Eye Movements and Reading Comprehension while Listening to Preferred and Non-preferred Study Music. *Psychol. Music* 40 (3), 339–356. doi:10.1177/0305735610387777

Johnson-Laird, P. N. (1993). "La théorie des modèles mentaux," in *Les modèles mentaux: Approche cognitive des représentations*.

Johnson-Laird, P. N. (1980). Mentel Models in Cognitive Science. *Cognitive Sci. A Multidiscip. J.* 4. doi:10.1207/s15516709cog0401_4

Jungnickel, E., and Gramann, K. (2016). Mobile Brain/Body Imaging (MoBI) of Physical Interaction with Dynamically Moving Objects. *Front. Hum. Neurosci.* 10, 306. doi:10.3389/fnhum.2016.00306

Kahneman, D. (1973). *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall.

Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition*. Cambridge: Cambridge University Press.

Kintsch, W., and Mangalath, P. (2011). The Construction of Meaning. *Top. Cognitive Sci.* 3 (2), 346–370. doi:10.1111/j.1756-8765.2010.01107.x

Kintsch, W. (1988). The Role of Knowledge in Discourse Comprehension: A Construction-Integration Model. *Psychol. Rev.* 95 (2), 163–182. doi:10.1037/0033-295X.95.2.163

Kintsch, W., and Van Dijk, T. A. (1978). Toward a Model of Text Comprehension and Production. *Psychol. Rev.* 85 (5). doi:10.1037/0033-295x.85.5.363

Lorch, R. F., Jr, Lorch, E. P., and Morgan, A. M. (1987). Task Effects and Individual Differences in On-Line Processing of the Topic Structure of a Text. *Discourse Process.* 10, 63–80.

Makransky, G., Mayer, R. E., Veitch, N., Hood, M., Christensen, K. B., Gadegaard, H., et al. (2019a). Equivalence of Using a Desktop Virtual Reality Science Simulation at Home and in Class. *PLoS One* 14 (4), e0214944. doi:10.1371/journal.pone.0214944

Makransky, G., Petersen, G. B., and Klingenberg, S. (2020). Can an Immersive Virtual Reality Simulation Increase Students' Interest and Career Aspirations in Science? *Br. J. Educ. Technol.* 51 (6), 2079–2097. doi:10.1111/bjet.12954

Makransky, G., and Petersen, G. B. (2021). The Cognitive Affective Model of Immersive Learning (CAMIL): a Theoretical Research-Based Model of Learning in Immersive Virtual Reality. *Educ. Psychol. Rev.* 33 (3), 937–958. doi:10.1007/s10648-020-09586-2

Makransky, G., Terkildsen, T. S., and Mayer, R. E. (2019b). Adding Immersive Virtual Reality to a Science Lab Simulation Causes More Presence but Less Learning. *Learn. Instr.* 60, 225–236. doi:10.1016/j.learninstruc.2017.12.007

Mayer, R. E. (2014). *Computer Games for Learning: An Evidence-Based Approach*. Cambridge: The MIT Press.

Mayer, R. E., Fiorella, L., and Stull, A. (2020). Five Ways to Increase the Effectiveness of Instructional Video. *Educ. Tech Res. Dev* 68 (3), 837–852. doi:10.1007/s11423-020-09749-6

Meyerhoff, H. S., and Huff, M. (2016). Semantic Congruency but Not Temporal Synchrony Enhances Long-Term Memory Performance for Audio-Visual Scenes. *Mem. Cogn.* 44, 390–390402. doi:10.3758/s13421-015-0575-6

Millis, K. K., Golding, J. M., and Barker, G. (1995). Causal Connectives Increase Inference Generation. *Discourse Process.* 20, 29–49. doi:10.1080/01638539509544930

Petersen, G. B., Petkakis, G., and Makransky, G. (2022). A Study of How Immersion and Interactivity Drive VR Learning. *Comput. Educ.* 179, 104429. doi:10.1016/j.compedu.2021.104429

Rayner, K. (1998). Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychol. Bull.* 124 (3), 372–422. doi:10.1037/0033-2909.124.3.372

Senkowski, D., Schneider, T. R., Foxe, J. J., and Engel, A. K. (2008). Crossmodal Binding through Neural Coherence: Implications for Multisensory Processing. *Trends Neurosci.* 31 (8), 401–409. doi:10.1016/j.tins.2008.05.002

Van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., and Theeuwes, J. (2008). Pip and Pop: Nonspatial Auditory Signals Improve Spatial Visual Search. *J. Exp. Psychol. Hum. Percept. Perform.* 34 (5), 1053–1065. doi:10.1037/0096-1523.34.5.1053

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.