

Implicit or Explicit: Gaze-Based Audio Commentaries for Ancient Painting Exhibitions in VR

Xin Ge, Xiaojiao Chen, Xiaoteng Tang, and Xiaosong Wang

QUERY SHEET

This page lists questions we have about your paper. The numbers displayed at left are hyperlinked to the location of the query in the paper.

The title and author names are listed here as they will appear in your paper and the Table of Contents. Please check that this information is correct and let us know if any changes are needed. Also check that affiliations, funding information and conflict of interest statements are correct.

Please review your paper as a whole for typos and essential corrections. Note that we cannot accept a revised manuscript at this stage of production or major corrections, which would require Editorial review and delay publication.

AUTHOR QUERIES

- Q1** The abstract is currently too long. Please edit the abstract down to no more than 150 words.
- Q2** Please provide volume, issue number and page range for the Ref. “Bauer-Krösbacher et al. 2013” and resupply if this is inaccurate.
- Q3** Please provide volume and issue number for the Ref. “Chen et al. 2024” and resupply if this is inaccurate.
- Q4** Please provide volume and issue number for the Ref. “Clemenson et al. 2020” and resupply if this is inaccurate.
- Q5** Please provide issue number for the Ref. “Cristina & Camilleri, 2018” and resupply if this is inaccurate.
- Q6** Please provide issue and doi number for the Ref. “Jacob 1993” and resupply if this is inaccurate.
- Q7** Please provide doi number for the Ref. “Jarrier & Bourgeon-Renault 2012” and resupply if this is inaccurate.
- Q8** Please provide missing publisher name for the "Jiménez-Hurtado & Soler Gallego 2015" references list entry.
- Q9** Please provide volume, issue number and page range for the Ref. “Kang et al. 2008” and resupply if this is inaccurate.
- Q10** Please provide issue number for the Ref. “Krukar & Dalton, 2020” and resupply if this is inaccurate.
- Q11** Please provide volume and issue number for the Ref. “Oppermann et al. 1999” and resupply if this is inaccurate.
- Q12** Please provide issue number for the Ref. “Quian Quiroga & Pedreira, 2011” and resupply if this is inaccurate.
- Q13** Please provide doi number and page range for the Ref. “Saito 2023” and resupply if this is inaccurate.
- Q14** Please provide volume, issue number and page range for the Ref. “Steffen et al. 2017” and resupply if this is inaccurate.
- Q15** Please provide issue number for the Ref. “Zhou et al. 2022” and resupply if this is inaccurate.
- Q16** Please note that the ORCID section has been created from information supplied with your manuscript submission/CATS. Please correct if this is inaccurate.



Implicit or Explicit: Gaze-Based Audio Commentaries for Ancient Painting Exhibitions in VR



Xin Ge , Xiaojiao Chen, Xiaoteng Tang , and Xiaosong Wang

Zhejiang University, Hangzhou, China

ABSTRACT

In modern museums that emphasize interactive and personalized experiences, classic audio commentaries are increasingly falling short of meeting visitors' expectations. Eye-tracking technology offers intuitive indoor pointing capabilities, presenting the potential to enhance visitor engagement and enable more personalized audio commentaries. However, previous research has focused on the technical aspects of gaze-based audio commentary systems, and research on paradigm and evaluation of eye-controlled interaction in gaze-based audio commentaries is still limited. In this study, we proposed three methods of gaze-based audio commentaries based on the implicit observations and direct control from the eye-controlled interaction paradigms and developed three prototypes: implicit, explicit, and implicit & explicit. Results from a within-subjects controlled study ($N=45$) with a baseline condition of no eye-tracking in virtual museums exhibiting ancient paintings indicated that the explicit method obtained better performance in terms of commentary experience, commentary quality, and intrinsic motivation, which were largely long-term. All three methods of gaze-based audio commentaries somewhat hindered users' acquisition of audio information, but they facilitated the acquisition of visual information. Among them, the implicit method supported users in learning the details of visual information, while the explicit method aided in learning the position and dimension of visual information. Nevertheless, they lacked long-term value for users' retention of the information. This study clarifies the design insights and strategies for implicit, explicit, and implicit & explicit methods of gaze-based audio commentaries, contributing to further exploring the potential of eye-controlled interaction in audio commentaries.

KEYWORDS

Eye-controlled interaction; audio commentary; user experience; cultural heritage; ancient painting

1. Introduction

With the popularization of information and communication technologies (ICTs) in museums (Machidon et al., 2018; Pujol-Tost, 2011; Rey & Casado-Neira, 2013), audio commentaries are gradually replacing on-site heritage interpreters as a primary means of information communication (Rich, 2016; Salmouka & Gazi, 2021; Waern et al., 2022). However, most classic audio commentaries follow a linear sequence of content playback, especially for explanations of individual artworks. While their content and order are carefully curated to offer a meaningful experience and learning, this rigid sequence cannot adapt to visitors' individual behaviors and motivations. This contradicts modern museums aiming to enhance visitor engagement and create more tailored experiences (Black, 2012; Parry, 2013).

In recent years, the rapid pace of technological advancement has created opportunities to enhance the personalization of museum audio commentaries (Ardissono et al., 2012; Lee, 2017; Sun & Yu, 2019). Some researchers have explored the integration of sensors into audio commentaries to capture visitors' interests and dynamically provide explanations aligned with their interests (Kaghat et al., 2020; Saito et al., 2023; Seidenari et al., 2017). These audio commentaries with

sensors are believed to offer visitors personalized explanations, enhancing the museum experience (Jarrier & Bourgeon-Renault, 2012; Pallud & Monod, 2010) and supporting cultural heritage learning (Sylaiou & Papaioannou, 2019; Yilmaz et al., 2024). Among them, eye-tracking technology as a sensor supporting eye-controlled interaction (Clay et al., 2019; Jacob, 1993; Majaranta & Bulling, 2014) has also been integrated into audio commentary systems to develop gaze-based audio commentaries (Dondi & Porta, 2023).

There are two types of gaze-based audio commentaries, which are rooted in the implicit observations and direct control from the eye-controlled interaction paradigms (Kumar et al., 2016; Majaranta & Bulling, 2014; Paulin Hansen et al., 1995). On the one hand, some audio commentary systems understand the areas of interest (AOIs) of visitors through gaze data and dynamically provide them with audio explanations (Mokatren & Kuflik, 2016; Raptis et al., 2021; Steffen et al., 2017; Toyama et al., 2011). On the other hand, there are audio commentary systems that enable visitors to directly control the playback of the audio explanations with their gazes (Mokatren et al., 2018; Piening et al., 2021). While each type has demonstrated potential in enhancing museum audio commentaries, there is still a lack of research examining the

eye-controlled interaction of gaze-based audio commentaries from a paradigm perspective.

Ancient paintings, as representatives of cultural heritage (Wollheim, 2023), are closely connected to the research field of audio commentaries in museums (Rashed et al., 2015; Saito et al., 2023; Vallez et al., 2020). However, the majority of audio commentaries are designed to provide explanations for all paintings within an entire gallery to visitors, which cannot enable them to control the explanations within paintings. This limits visitors' opportunity to receive personalized audio commentaries tailored to the individual painting. In recent years, some studies have produced gaze-based audio commentaries for individual ancient paintings based on eye-tracking technology (Raptis et al., 2021; Yang & Chan, 2019; Zimmermann & Lorenz, 2008). They divided the painting into distinct active areas, each linked to specific audio commentary that could be triggered by the visitor's gaze. However, their research goal is limited to the technical aspect, and the values of the experience and learning have yet to be assessed.

In this study, we proposed the methods of "implicit" and "explicit" gaze-based audio commentaries, which were derived from the eye-controlled interaction paradigms. The implicit method captured users' focused AOIs by analyzing the gaze data in real-time and provided the associated audio commentaries (Figure 1(a)). The explicit method allowed users to activate the translucent red speaker icons with their gazes to directly control the playback of audio commentaries (Figure 1(b)). Recognizing the limitations of each method, such as the lack of control in the implicit method and the disruption of vision in the explicit method, and their potential complementary strengths, we introduced "implicit & explicit" as the third method for gaze-based audio commentaries. The implicit & explicit method involved two steps: first, capturing the user-focused AOI, and then providing a translucent red speaker icon for the user to control the playback of the audio commentary (Figure 1(c)).

With the integration of eye-tracking technology into virtual reality head-mounted displays (VR HMDs) (Plopski et al., 2022), the creation of eye-controlled interactions in virtual museums using eye-tracking technology has become increasingly feasible. Therefore, we employed a VR HMD with eye-tracking technology, based on three methods of gaze-based audio commentaries, with a baseline condition of no eye-tracking (classic audio commentary), to develop four virtual museums corresponding to the four conditions: no eye-tracking (N), implicit (I), explicit (E), and implicit & explicit (IE). This study used virtual museums exhibiting traditional Chinese landscape paintings from the Song Dynasty as design case studies (Figure 1(d)).

To understand the differences between the three methods of gaze-based audio commentaries in visitors' experience and learning while appreciating ancient paintings, three research questions were posed in this study.

RQ1: Do the three methods of gaze-based audio commentaries enhance the experience and quality of audio commentaries?

RQ2: Do the three methods of gaze-based audio commentaries improve users' acquisition of audio information and visual information?

RQ3: Do the three methods of gaze-based audio commentaries enhance intrinsic motivation and reduce cognitive load?

To answer the three research questions, we employed a mixed-methods approach and a within-subjects design in a controlled study to examine the effects, including commentary experience, commentary quality, audio information learning, visual information learning, intrinsic motivation, and cognitive load. The specific comparison metrics of the commentary experience were engagement, knowledge/learning, meaningful experience, and emotional connection. The specific comparison metrics of the commentary quality were general usability, learnability and control, and quality of interaction with the guide. The specific comparison metrics of the audio information learning were *meaning of artwork*,

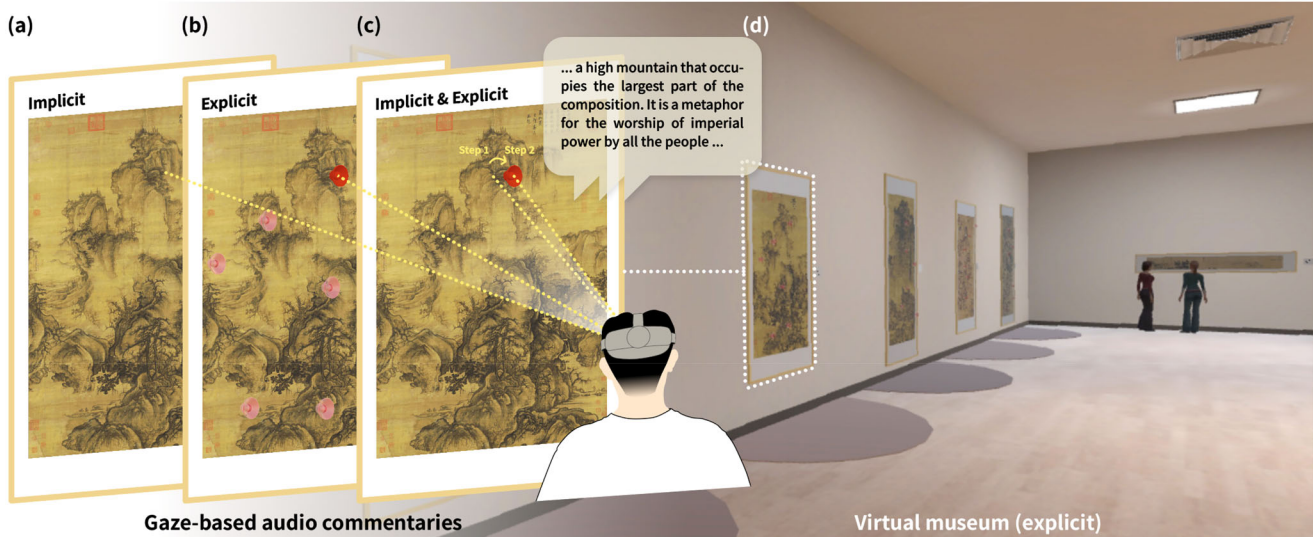


Figure 1. A visitor wearing a VR HMD enters a virtual museum environment to appreciate ancient paintings. (a) The visitor is using the implicit method of gaze-based audio commentaries. (b) The visitor is using the explicit method of gaze-based audio commentaries. (c) The visitor is using the implicit & explicit method of gaze-based audio commentaries. (d) An exhibition showcasing traditional Chinese landscape paintings from the Song Dynasty, which uses the explicit method of gaze-based audio commentaries.

motivation of artwork, production process, opinions of artist, and artist information. The specific comparison metrics of the visual information learning were *position, dimension, and detail.* The last two aspects (intrinsic motivation and cognitive load) offered further explanations for both experience and learning. Additionally, data collection took place at two-time points: immediately following the experience (time 1, short-term assessment) and approximately one month later (time 2, long-term assessment). We also conducted a qualitative analysis to enrich the discussion, using semi-structured interviews to discuss the advantages and disadvantages of the four conditions. Experimental results showed the positive impact of the three methods of gaze-based audio commentaries on visitors' experience and learning to a certain extent, as well as their respective advantages and disadvantages, thereby enabling the identification of design insights and strategies for implementing them in museums.

Our research makes three main contributions. (1) After a discussion of the eye-controlled interaction paradigms, three methods of gaze-based audio commentaries were proposed, and their parameters were also determined. (2) We conducted a controlled study in VR environments to assess their effects on experience and learning and get design insights. (3) By integrating suggestions derived from participants, we identified improved strategies to enhance the design of gaze-based audio commentaries in the future.

2. Related work

2.1. Personalized audio commentaries in museums

The concepts of audio commentaries and audio guides are similar, as they both offer audio explanations of artworks in museums (Jelavić et al., 2012; Saito, 2023; Saito et al., 2023). Some audio guides also direct visitors to specific artworks while offering audio explanations (Lee, 2017). Therefore, there is a notable overlap between audio commentaries and audio guides in research related to museums. In this study, since the primary function of our system is to explain the background and content of paintings, we adopt the concept of audio commentaries uniformly.

Personalization refers to a system's ability to observe user behavior, make decisions and predictions based on their interests, and deliver content that is tailored accordingly (Fan & Poole, 2006). In museums, personalized information systems, including audio commentaries, should be responsive to the prevailing context and the attention of the visitors in order to provide information that aligns with their expectations. Research aimed at achieving this goal is primarily divided into two major categories. On the one hand, some researchers have explored the use of dynamic modeling as a means of analyzing visitors' interests. For example, Oppermann et al. (1999) introduced a nomadic information system "Hippie," which created contextualized information spaces based on visitors' interests, knowledge, and preferences to provide personalized information services. On the other hand, some studies have considered information systems with sensors to perceive and respond to visitors' interests in real-time. For example, Kaghat and Cubaud (2010)

developed a location-aware audio guide device called "SARIM." It sensed the position and orientation of the visitor's head and provided an adaptive museum information experience. However, existing studies mainly depend on indirect data, such as visitors' previous data and physical positioning, to deduce their characteristics and interests. This limitation fails to satisfy the increasing demand for more detailed and personalized information services in museums.

In recent years, eye-tracking as a sensing technology has been shown to enable intuitive indoor pointing in museums (Mokatren et al., 2018; Mokatren & Kuflik, 2016). Its use enables information systems to determine visitors' AOIs with greater precision by analyzing the position of their gaze points, providing more accurate personalized information services. Some researchers have developed personalized audio commentary systems that employ eye-tracking technology. For example, Toyama et al. (2011) developed a personalized audio guide system called "Museum Guide 2.0," which used a head-mounted mobile eye tracker prototype to deliver audio commentaries to visitors about the specific artworks they were viewing. In addition, several studies have explored eye-tracking technology to provide personalized audio commentaries of visitors' AOIs within individual artworks. For example, Raptis et al. (2021) created an eye-controlled interface for interacting with a single painting, where the user could move the gaze point to AOIs in the painting to get audio commentaries of the presentations. Furthermore, based on facilitating the personalized information service, eye-tracking technology further helps to enhance the positive impact of audio commentaries on the museum experience and learning. For example, Mokatren et al. (2018) compared an audio guide system integrated with a mobile eye tracker in a small museum and found that participants recognized this personalized interaction system's contribution to the visit experience, especially in improving the learning outcome.

Overall, these preliminary studies of gaze-based audio commentaries demonstrate that eye-tracking technology offers novel opportunities to enhance the personalized information services of audio commentaries. However, research on the usage of eye-tracking technology in museum audio commentaries is still rare, and further design and evaluation of gaze-based audio commentaries are currently needed.

2.2. Eye-controlled interaction in museums

Eye-controlled interaction is a novel method of using eye-tracking technology to enable users to operate devices via their gazes (Cristina & Camilleri, 2018; Hou & Chen, 2021; Ruhland et al., 2015). In recent years, eye-controlled interaction has been initially explored and researched in the field of cultural heritage, with the results representing the latest developments in museum interactive applications (Dondi & Porta, 2023). Following the implicit observations and direct control from the eye-controlled interaction paradigms (Kumar et al., 2016; Majaranta & Bulling, 2014; Paulin Hansen et al., 1995), we categorized the eye-controlled interactions in these studies into "implicit" and "explicit" categories.

Regarding implicit eye-controlled interaction in museums. In the early stages, a visual representation of the gaze data was used to facilitate interaction between the visitor and the artwork. For example, Wooding et al. (2002) conducted eye-tracking experiments to capture visitors' gazes at paintings in real-time during an exhibition, simultaneously displaying each visitor's gaze on a large public screen, enabling a form of interactive engagement. In addition, ARoS Art Museum,¹ Cleveland Museum of Art,² and M – Museum Leuven³ incorporated eye-tracking technology into their exhibitions, allowing participants and other visitors to observe where the gazes move while viewing artworks, and even providing comparisons of the differences in focus between visitors. Although this approach proved effective in engaging visitors, it did not extend to the exploration of the potential of implicit eye-controlled interaction as a means of interacting with artwork information. In recent years, some studies have captured visitors' interests by analyzing their gaze data and dynamically providing information about the artworks they were interested in. For example, Zimmermann and Lorenz (2008) created the LISTEN system aiming to hide technology within the museum environment. It allowed the audio that visitors heard to depend on their movement and gaze direction within the museum. Yang and Chan (2019) provided spatialized sounds for individual paintings based on visitors' visual behavior through eye-controlled interaction.

Regarding explicit eye-controlled interaction in museums. It is prioritized in museums' accessibility strategies, as it can serve as an alternative to touch for people with disabilities or in situations where there is a risk of epidemiological transmission. For example, Dondi et al. (2022) developed an eye-controlled application that enabled visitors to directly control multimedia content in art images with their gazes. This enabled people with disabilities to interact with the artworks on display. In addition, some researchers have explored the use of explicit eye-controlled interaction to provide information for general visitors. For example, Al-Thani and Liginlal (2018) designed explicit eye-controlled interaction in a virtual gallery based on eye-tracking technology. Visitors could directly control the browsing and exploration of photos, songs, and videos related to the artwork through gazing. Our previous study introduced the concept of "actively viewing audio-visual media" through explicit eye-controlled interaction, enabling visitors to actively control the content of audio-visual media by directing their gazes (Chen et al., 2024).

Previous studies have demonstrated that the "implicit" and "explicit" categories of eye-controlled interaction both have rich application scenarios in museums. However, their respective advantages and disadvantages in museum environments remain unclear. Therefore, further research is needed to assess the values of eye-controlled interaction in museum environments from a paradigm perspective.

2.3. VR interactive applications of painting exhibitions

The use of immersive technologies can facilitate an immersive museum experience for visitors, thereby enhancing the

accessibility of cultural heritage content (Capece et al., 2024; Innocente et al., 2023; Yi & Kim, 2021; Zhou et al., 2022). There are a number of studies conducted with the objective of incorporating VR into interactive applications of painting exhibitions in order to enhance visitors' experience and learning of paintings. Virtual reconstruction and roaming are the most common VR interactive applications for creating painting experiences. For example, Yuan and Yun (2016) reconstructed the virtual scenes and characters depicted in the painting *Listening to the Qin* (听琴图) based on a VR HMD. This approach permitted viewers to interact with the characters and become immersed in the painting. Additionally, some studies have developed virtual galleries accessible via VR to create ideal viewing experiences of paintings. For example, Hürst et al. (2016) not only recreated art galleries in VR environments but also digitally extended the representation of paintings by incorporating style migrations and adding 3D animations, which were found to have a positive impact on the viewing experiences of paintings. Beyond their experiential value, some studies have also explored the potential value of VR interactive applications for improving visitors' aesthetic and philosophical understanding of paintings. For example, Kuo et al. (2024) reconstructed a modern painting as an immersive VR experience, enabling students to enter the painting for art appreciation. They found that this approach enhanced students' understanding and learning of the painting's concepts. Jin et al. (2022) reconstructed a freely explorable VR painting space for a traditional Chinese painting *Spring Morning in the Han Palace* (汉宫春晓图), which allowed collaborative learning and proved to yield good learning outcomes.

The above research suggests that VR interactive applications have been initially studied in painting exhibitions and show potential for improving the experience and learning of such artworks. However, there has been limited research on audio commentaries of painting exhibitions in VR, particularly those incorporating novel human-computer interaction techniques like eye-tracking, which remain largely unexplored.

3. Prototype design

3.1. Design of gaze-based audio commentaries

3.1.1. Concept of the three methods of gaze-based audio commentaries

There are two major paradigms for eye-controlled interaction by eye-tracking: implicit observations and direct control (Kumar et al., 2016; Majaranta & Bulling, 2014; Paulin Hansen et al., 1995). Implicit observations predict user intent based on eye movement signals, which can be leveraged to enhance viewing activities. In museum environments, this can be used for audio enhancement, enriching the visitor's sensory experience. Direct control involves the user deliberately inputting interaction commands to the system by moving eyes. In museum environments, it enables lightweight interactions, such as simple selections and confirmation operations. However, implicit observations and direct control each have their drawbacks. The integration of the two is thought to overcome the shortcomings of lack of

control due to the system's reliance on implicit observations, as well as the issue of disrupted vision caused by using only direct control for eye-controlled interactions (Kumar et al., 2016). Accordingly, an integration of the two represents one of the key eye-controlled interaction paradigms and is considered a better option.

The incorporation of eye-controlled interaction into classic audio commentaries addresses the limitation of audio sequential fixation. Furthermore, the natural and intuitive advantages of this interaction facilitate significant enhancements to the museum experience and learning (Mokatren et al., 2018; Raptis et al., 2021). This study aims to explore the potential of eye-controlled interaction in audio commentaries from a paradigm perspective. We designed the three methods of gaze-based audio commentaries for exhibitions of traditional Chinese landscape painting, based on the eye-controlled interaction paradigms. In addition, we also included the classic audio commentary as a baseline condition, resulting in a total of four conditions and corresponding prototypes: no eye-tracking (N), implicit (I), explicit (E), and implicit & explicit (IE). Table 1 presents the interactive interfaces and interaction logic in the four conditions.

3.1.2. Design details of eye-controlled interfaces

In the design of the eye-controlled interface in condition I, we defined the AOIs on the painting based on the content of the audio commentaries, linking them to their respective audio commentaries. We used the dispersion-threshold identification (I-DT) fixation algorithm (Salvucci & Goldberg, 2000), as suggested by Kwok et al. (2019), to obtain the location of users' attention in real-time and provide audio commentaries associated with AOIs they focused. The I-DT fixation algorithm is suitable for fixed observers, which requires two parameters, including dispersion threshold and duration threshold. The acceptable range for the dispersion threshold is between 0.7° and 1.6° of visual angle (Blignaut, 2009; Llanes-Jurado et al., 2020), and the duration threshold is typically set to a value between 100 ms and 400 ms (Llanes-Jurado et al., 2020; Widdel, 1984). However, in a museum setting where accurately triggering audio commentaries based on visitor interest or intent is essential, the I-DT fixation algorithm needs to be further evaluated to determine the optimal parameters.

In the design of the eye-controlled interface in condition E, we used dwell time (Jacob, 1990) as an explicit interaction method to activate the static buttons based on previous research (Al-Thani & Liginlal, 2018; Chen et al., 2024; Piening et al., 2021). The static buttons were represented by translucent red speaker icons (3D) linked to the audio commentaries. They were placed above the visual elements of the painting to indicate locations where audio commentaries could be triggered. The process of dwell time was visualized through a fill animation (Figure 2). Specifically, when the line of sight intersects with the translucent red speaker icon, an opaque red speaker icon begins to appear inside the translucent red speaker icon, gradually filling it. The audio commentary is triggered once the translucent red speaker icon is fully filled after a certain dwell time. To minimize


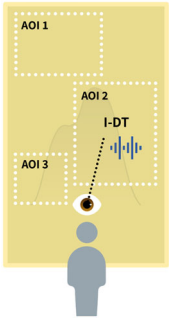
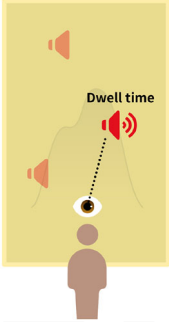
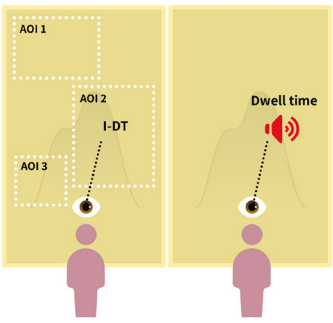
the issue of speaker icons disrupting the vision, we referred to a previous study to minimize their dimensions, which suggested that the minimum dimensions for static buttons in eye-controlled interfaces should range from 2.636° to 3.322° of visual angle (Niu et al., 2021). In addition, the speaker icons were designed to disappear immediately upon activation and reappear only after the audio commentaries had finished playing. To avoid the Midas touch problem, some researchers have implemented dwell times between 150 ms and 1000 ms (Helmert et al., 2008; Jacob, 1990, 1991; Magee et al., 2015). However, there are no established parameters of button dimension and dwell time specifically suited for gaze-based audio commentaries in museum environments, necessitating further estimation.

In the design of the eye-controlled interface in condition IE, the eye-controlled interface in the first step follows the design of condition I. After the first step, a translucent red speaker icon appears at the center of the user's focused AOI, leading to the second step. The eye-controlled interface in this step follows the design of condition E. After the second step is completed, the speaker icon is activated, and its associated audio commentary is triggered. Overall, condition IE forms a combined eye-controlled interface by merging the designs of conditions I and E. Thus, this eye-controlled interface is relatively more complex and requires the parameters from both conditions I and E (I-DT and dwell time).

3.1.3. Pre-study: Determining optimal parameters for eye-controlled interfaces

We pre-studied 10 participants in virtual museums to determine the optimal parameters for the eye-controlled interfaces of gaze-based audio commentaries in the two methods (I-DT and dwell time). Based on parameter ranges and levels by previous research, we set three levels for the dispersion threshold (0.7° (18 mm), 1° (26 mm), and 1.6° (42 mm)), using the distance dispersion algorithm (Shic et al., 2008) to define the dispersion among gaze points, the distance between user and painting = 1.5 m) and duration threshold (100 ms, 250 ms, and 400 ms) in the I-DT method, and three levels for the button dimension (2.636° (69 mm), 2.978° (78 mm), and 3.322° (87 mm)), expressed as the length of each side of the cube that holds the 3D speaker icon, the distance between user and painting = 1.5 m) and dwell time (150 ms, 600 ms, and 1000 ms) in the dwell time method. Two experiments were conducted afterward. In the first experiment, we developed 3 (dispersion threshold) \times 3 (duration threshold) = 9 eye-controlled interfaces of gaze-based audio commentaries for the I-DT method. In the second experiment, we developed 3 (button dimension) \times 3 (dwell time) = 9 eye-controlled interfaces of gaze-based audio commentaries for the dwell time method. In each experiment, participants were asked to use each eye-controlled interface individually within the corresponding virtual museum and evaluate usability using a 7-point Likert scale, including the four indicators of effectiveness, efficiency, satisfaction, and ease. Then, we identified the eye-controlled interface with the highest mean score in each experiment and used its parameters as the optimal

Table 1. Interactive interfaces, interaction logic, and optimal parameters in the four conditions.

Condition	Interactive interface	Interaction logic	Optimal parameter			
			I-DT		Dwell Time	
			Dispersion Threshold (°)	Duration Threshold (ms)	Button Dimension (°)	Dwell Time (ms)
N: no eye-tracking (baseline)		The system calculates the distance between the user and each painting in real-time. When the user moves closer to a painting (distance ≤ 1.5 m), the audio commentary is triggered, which is looped until the user leaves.	–	–	–	–
I: implicit		The painting is divided into several invisible AOIs based on the content of the audio commentaries. The system detects in real-time whether the user is focusing on a specific AOI (i.e., detection of gaze fixation), and if so, it triggers the audio commentary associated with that area.	1	100	–	–
E: explicit		Several translucent red speaker icons, representing static buttons, are above the painting. These speaker icons are connected to the audio commentaries and indicate the audio commentaries can be triggered at those locations. When the system detects that a user is focusing on one of the speaker icons for a certain dwell time, the audio commentary associated with that speaker icon is triggered.	–	–	3.322	600
IE: implicit & explicit		The process involves two steps. First, when the system detects a gaze fixation within an AOI, it displays the translucent red speaker icon associated with that AOI. Then, the user can choose to trigger the associated audio commentary by focusing on the speaker icon for a certain dwell time.	1	100	3.322	600

parameters for the eye-controlled interface in each method, as shown in Table 1.

3.2. Painting selection

The ancient paintings incorporated into audio commentaries were traditional Chinese landscape paintings with extensive

intersections with digital media (Cheng, 2024). They were selected by an expert panel (five teachers with backgrounds in art and archaeology) based on criteria including knowledge quantity, informational composition, and cognitive difficulty. (1) The selected paintings are typical of traditional Chinese landscape paintings featuring mountains, trees, buildings, and people. These paintings contain numerous visual elements

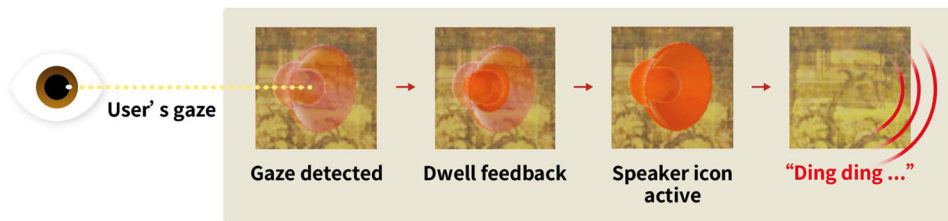


Figure 2. The eye-controlled interface of the dwell time method from the user's perspective. When the user gazes at the translucent red speaker icon, an animation fills it, providing feedback on the progress of the dwell time. Once the translucent red speaker icon is fully filled, it is activated, emitting a "ding ding" sound and playing the audio commentary.

accompanied by detailed explanatory texts, making them ideal for providing content for audio commentaries. (2) The selected paintings have comparable original dimensions, with a height approximately equal to that of an adult. Additionally, each painting features a mountain as the central visual focus, with trees, buildings, and people arranged around it. Their large size and loose composition facilitate the delineation of the AOIs. (3) The selected paintings were created during the early Song Dynasty, reflecting the historical and cultural context of that period. They both convey a sense of majesty through the expansive composition and thus share a similar aesthetic foundation and cognitive difficulty. In the end, we chose four traditional Chinese landscape paintings to be used in the four conditions (Figure 3): Guo Xi's *Early Spring* (早春图),⁴ Fan Kuan's *Travelers Among Mountains and Streams* (溪山行旅图),⁵ Jing Hao's *Mount Kuanglu* (匡庐图),⁶ and Dong Yuan's *Along the Riverbank* (溪岸图).⁷

3.3. Audio content

The role of museum interpretation is to guide visitors in exploring collections by providing key information about the artwork, the artist, and the cultural context, as deemed significant by curators (Serota, 1997). This indicates that identifying the key information about the paintings is essential for designing the audio content of audio commentaries. To achieve this, we first gathered information about the four selected paintings from the official websites of the museums where they were housed, as well as from Google Arts & Culture and Wikipedia. Afterward, we classified the information on the paintings into five categories, following the categorization of artwork information elements from Yi et al. (2021): "(1) meaning of the artwork, (2) motivation behind the artwork, (3) production process, (4) artist's opinions, and (5) artist's information." Since painting interpretations focus on visual information, visitors listening to audio commentaries rely heavily on descriptions of the visual elements in the painting. To address this, we incorporated visual descriptions into the audio content and proposed a composition structure of "visual description + painting information." Specifically, each audio content begins with a description of a visual element in the painting, followed by related information from the five artwork information elements mentioned above. Finally, the expert panel discussed and selected the six most important pieces of information for each painting to be used as audio content.

In order to reduce audio content cognitive load and enhance memorability (Kim et al., 2008), we referred to audio descriptive guides (ADG) (Jiménez-Hurtado & Soler Gallego, 2015) to add spatial positioning cues to the audio content. These cues were placed at the beginning or ending of each audio content. However, due to differences in interaction methods across the four conditions, their spatial positioning cues in audio content varied. Specifically, condition N required spatial positioning cues at the beginning of each audio content to indicate the location of the visual element associated with the audio content on the painting. In contrast, the three methods of gaze-based audio commentaries did not require this, as eye-controlled interaction ensured the audio content corresponded to the visual elements the user focused on. For conditions I and IE, additional spatial positioning cues for nearby AOIs were added at the end of each audio content to guide the users' gazes, as users were unsure where the next audio commentaries could be triggered. For condition E, the speaker icons acted as spatial positioning cues, so no additional cues were needed within the audio content. Table 2 lists two examples of audio content across the four conditions.

All the audio content of the audio commentaries was recorded and voiced by a text-to-speech (TTS) tool,⁸ using an upbeat female tone to enhance memorization (Velentza et al., 2020). To control the experimental variables, these audio commentaries were made to have similar lengths.

4. Evaluation study

4.1. Study design

The study received explicit ethics approval from the Zhejiang University Ethics Committee and was conducted in a controlled laboratory environment at Zhejiang University. All data collected during the experiment were anonymized to ensure participant confidentiality. Eye-tracking data were exclusively used for real-time eye-controlled interaction within the experimental framework and were neither recorded nor stored post-experiment.

We built four virtual museums corresponding to the four conditions, each containing the same set of traditional Chinese landscape paintings including the four selected paintings for the experiment. Participants were able to appreciate these ancient paintings in four conditions using a VR HMD: PICO 4 Pro. It adds two eye-tracking infrared cameras on the inside of the HMD, which provides accurate and stable eye-tracking.



Figure 3. The four selected paintings were located on the same exhibition wall in the virtual museums. They were arranged in a fixed order from left to right.

Table 2. Two examples of audio content across the four conditions.

Example Number	Condition	Beginning	Visual Description	Painting Information	Ending
Ex.1	N	Above it is	a high mountain that occupies the largest part of the composition.	It is a metaphor for the worship of imperial power by all the people. (...) It is also one of the most important artworks in the entire history of Chinese art.	–
	I	You are looking at			You can continue to look left or down.
	E	This is			–
	IE	This is			You can continue to look left or down.
Ex. 2	N	The left margin is	an inscription by the artist Guo Xi, which states, "Painted by Guo Xi in the early spring of the year Renzi (壬子)."	Guo Xi worked at the Northern Song Painting Academy. (...) He is both a famous painter of the Northern Song Dynasty and a renowned theorist of painting.	–
	I	You are looking at			You can continue to look upper right or lower right.
	E	This is			–
	IE	This is			You can continue to look upper right or lower right.

The bolded texts are the spatial positioning cues in audio content.

4.2. Participants

The participants were college students with normal or corrected vision. They were tested on their knowledge of traditional Chinese landscape painting, and those with no or minimal knowledge were eligible to participate. We used G*Power software to determine the minimum sample size, as recommended by Faul et al. (2009). The effect size was set to medium (0.25) with an alpha level of 0.05 and a required power of 0.80. The calculation result indicated that 24 participants were needed for the experiment. In our experiment, the total number of participants was 47. One month later (time 2), we invited all participants to return for the follow-up questions of the study. Two participants withdrew, leaving us with 45 participants, 23 males and 22 females (age $M = 25.20$, $SD = 2.83$). Among these participants, 75.6% of them visited the museum more than once yearly (times $M = 3.00$, $SD = 2.05$). The participants were asked to indicate their level of agreement with a series of attitude statements on a 5-point Likert scale, where 1 represented "strongly disagree" and 5 represented "strongly agree." Regarding audio commentaries, 80.0% of the participants had used them before and had a positive attitude towards using them in museums ($M = 3.73$, $SD = 0.92$). Regarding eye-controlled interaction, 37.8% of the participants had experienced it before. Overall, they had limited familiarity with eye-controlled interaction ($M = 2.51$, $SD = 1.08$).

4.3. Measures

A mixed-methods approach and a within-subjects design were employed to investigate the effects of the four conditions,

incorporating both quantitative and qualitative methods. We began by collecting and analyzing quantitative data to address the three research questions in our study. Subsequently, we gathered qualitative data to explain the quantitative findings from the first stage, gaining deeper insights (Creswell & Clark, 2017; Ivankova et al., 2006). Descriptive analysis (mean (M) and standard deviation (SD)), the one-way ANOVA test, and the paired samples t-test were used to compare the differences in the mean-rated values and scores between conditions and between times. For all items in the questionnaires mentioned below, Cronbach's alpha test was used to measure the reliability of the scale.

To address RQ1 (commentary experience and commentary quality), we collected feedback on each participant's commentary experience and commentary quality in each condition, using the four-factor Museum Experience Scale (MES) and the three-factor Multimedia Guide Scale (MMGS) (Othman et al., 2011). The MES measured engagement, knowledge/learning, meaningful experience, and emotional connection. The MMGS measured general usability, learnability and control, and quality of interaction with the guide. We asked participants to rate the MES and MMGS on a 5-point Likert scale.

To study RQ2 (acquisition of audio information and visual information), we aimed to assess participants' learning outcomes in terms of the acquisition of audio information and visual information. Audio information referred to the content learned from the audio commentaries, while visual information referred to the content learned from the canvases of the ancient paintings. In terms of audio information, a pre-test and post-test experimental design was conducted, where

participants answered a test designed by the research team. The pre-test and post-test contained identical content, but the question order varied. Specifically, the test was designed to measure learning outcomes of audio information across five dimensions based on the five artwork information elements, including (1) *meaning of artwork*, (2) *motivation of artwork*, (3) *production process*, (4) *opinions of artist*, and (5) *artist information*. Each content dimension was examined by a multiple-choice question and a multiple-choice question (select one or more answer choices). Therefore, each condition had five multiple-choice questions and five multiple-choice questions (select one or more answer choices) for a total of 10 multiple-choice questions. Each multiple-choice question had four options and an "I don't know" option. The option "I don't know" was included to eliminate the estimation effect, which could compromise the reliability of multiple-choice tests (Chiu & Camilli, 2013; Zhang, 2013). To ensure the content validity of the test, it was recommended to gather expert reviews (Fraenkel & Wallen, 1990). Consequently, after the test was created, the expert panel was invited to review the questions. The average difficulty of the test across the four conditions ranged from 0.4 to 0.7 on a standardized scale (Ahmann & Glock, 1981). All questions exhibited item discrimination levels that exceeded the 0.2 standard (Ebel & Frisbie, 1991). In terms of visual information, participants were asked to recall as many visual elements of the original paintings as possible and draw sketch paintings within empty drawing canvases on a tablet. If participants found it difficult to accurately draw the visual elements from memory, they were allowed to describe them in words and write their descriptions on the canvas. We measured learning outcomes of visual information by sketch ratings, which were assessed along three dimensions: (1) *position*, (2) *dimension*, and (3) *detail*. They were based on the visual principles of traditional Chinese landscape painting as described in *Linquan Gaozhi* (林泉高致)⁹ which was a Song Dynasty treatise on landscape painting. The score for each dimension was determined by the number of visual elements that met the criteria for that dimension. For instance, in the *position* dimension, visual elements placed correctly on the canvas would be coded, and their count would represent the score for the *position* dimension. Two researchers independently coded each sketch painting according to the rules, recording scores for the three dimensions as well as a total score. In addition, they recorded the proportion of scores for visual elements associated with the audio commentaries in each sketch painting (i.e., the percentage of audio-related scores in the total score). During the coding and scoring process, a double-blind method was employed to ensure the results were not influenced by bias. We conducted intra-class correlation (ICC) to assess the inter-rater reliability of the two coders. According to the 95% confidence interval for the ICC estimate, the following classification criteria were applied to evaluate reliability levels: "Values less than 0.5 are indicative of poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values greater than 0.90 indicate excellent reliability." (Koo & Li, 2016) The respective average measure ICC values for the *position*, *dimension*, and *detail* were 0.912 (Excellent), 0.916 (Excellent), and 0.895 (Good) at

time 1, and were 0.914 (Excellent), 0.917 (Excellent), and 0.879 (Good) at time 2.

To address RQ3 (*intrinsic motivation and cognitive load*), we collected feedback on each participant's intrinsic motivation and cognitive load in each condition, using Intrinsic Motivation Instrument (IMI) (Ryan & Deci, 2000a, 2000b) and NASA Task Load Index (NASA-TLX) (Hart & Staveland, 1988). The NASA-TLX questionnaire is a highly regarded tool for measuring cognitive load across diverse environments (Hart, 2006). The widespread application of NASA-TLX in numerous VR studies demonstrates its applicability within the VR environment (George et al., 2018; Reinhardt et al., 2019; Tabanfar et al., 2018). The IMI consisted of four factors, including interest/enjoyment, perceived competence, perceived choice, and pressure/tension. Participants were asked to rate the IMI using a 7-point Likert scale. The NASA-TLX had six items, including mental demand, physical demand, temporal demand, effort, performance, and frustration level, with values calculated on a scale ranging from 0 to 100.

Additionally, we conducted semi-structured interviews with participants to gain further insights into the three research questions explored in our study. The semi-structured interview included two construction questions: (1) Discuss the advantages and disadvantages of the four conditions regarding both the experience and learning. (2) Provide recommendations to enhance the experience and learning of gaze-based audio commentaries. During the interviews, we requested consent from participants to use an audio recorder for interview transcription data. After that, the interview transcription data were independently collated by two researchers, who identified themes. These themes were then discussed with a third researcher, and the summary themes were confirmed.

4.4. Procedures

At the beginning of each experiment, the experimenter provided the participants with an overview of the study. The experiment consisted of two parts: time 1 and time 2 (Figure 4). Time 1 took an average of 60 min. First, participants were asked to complete a personal and background survey, including demographic information and pre-test multiple-choice questions, as well as an informed consent form. After that, they were asked to wear the VR HMD and calibrate the eye-tracking. The first two steps lasted approximately five minutes. Second, participants entered the four virtual museums, corresponding to the four conditions in order of Latin square.¹⁰ Each time participants entered a virtual museum, the experimenter directed them to a painting that was not used in the actual study and described how to use the audio commentaries in that condition. Once participants reported being comfortable with the audio commentaries, they were directed to one of the four selected paintings in the experiment to use the audio commentaries. Only one painting was viewed in each condition, for a total of four selected paintings, and their viewed order was fixed. We did not provide participants with explicit goals for experience and learning. In addition, there was no time limit on how

long they could view the painting or listen to the audio commentaries. Following the completion of each condition, participants were required to complete a post-test questionnaire and a post-survey questionnaire. The post-test questionnaire, including multiple-choice questions and sketch paintings, assessed the acquisition of audio information and visual information. The post-survey questionnaire evaluated commentary experience (MES), commentary quality (MMGS), intrinsic motivation (IMI), and cognitive load (NASA-TLX). After that, participants were asked to move to the virtual museum of the next condition until all conditions were completed. This step lasted approximately 50 min. Subsequently, semi-structured interviews were conducted with the participants, with each interview session lasting approximately five minutes. At the end of the interview, we thanked the participants and invited them to return to the second part of the experiment at time 2. We also informed the participants not to intentionally view or study traditional Chinese landscape paintings before time 2. However, it is imperative to acknowledge the perspective of Falk and Dierking (2000), who argue that memory and learning are dynamic processes shaped by daily experiences. Consequently, our study allowed for naturally occurring learning behaviors, such as engaging in self-directed recall or drawing connections between elements of ancient paintings and events from their daily experiences. Approximately a month later, participants received the questionnaires for experiment time 2, which were identical to those for time 1.

We revised the wording of the questionnaire items to align with the specific context of this study. Before conducting the evaluation study, a pilot study with six participants was conducted to ensure that both task designs and questionnaire items were appropriate. The analysis of the pilot study data revealed generally positive results in terms of internal consistency.

5. Results

In this section, all quantitative data analyses were conducted using SPSS 26. Qualitative data analysis was conducted using NVivo for theme-based content analysis (Neale & Nichols, 2001). To enhance the credibility of the study, we sent the transcription of the taped interview to the interviewee for verification, ensuring consistency between the recorded and transcribed data (Creswell & Creswell, 2018). Based on the interviewees' responses, the data appeared to be consistent.

5.1. Quantitative results

5.1.1. Commentary experience and commentary quality

Figure 5(a and b) show the results of the MES. At time 1, the Cronbach's alpha test yielded a high level of internal consistency for all items of commentary experience (20 items, $N=0.860$, $I=0.888$, $E=0.880$, $IE=0.881$). There was a trend that condition E was rated highest in terms of all factors. Although condition I was rated lowest in terms of engagement and knowledge/learning. A one-way ANOVA test revealed that there were significant differences between conditions in terms of engagement ($F_{3,176}=4.327$, $p=0.006$) and knowledge/learning ($F_{3,176}=3.622$, $p=0.014$). Post-hoc tests

indicated significant differences between conditions I and E in terms of engagement ($p=0.003$) and knowledge/learning ($p=0.015$). The differences between conditions E and IE were significant in terms of knowledge/learning ($p=0.049$). At time 2, the Cronbach's alpha test yielded a high level of internal consistency across all items related to the commentary experience (20 items, $N=0.859$, $I=0.932$, $E=0.935$, $IE=0.933$). Condition E was still rated highest overall. A one-way ANOVA test revealed that there were significant differences between conditions in terms of knowledge/learning ($F_{3,176}=4.543$, $p=0.004$). Post-hoc tests indicated significant differences between conditions N and I ($p=0.050$) and between conditions I and E ($p=0.003$) in terms of knowledge/learning. Results of paired samples t-tests between times 1 and 2 in all factors of the MES are shown in Table 3. There were no significant differences in the values of most factors (4 factors, $N=4/4$, $I=3/4$, $E=3/4$, $IE=4/4$, X/Y: X out of Y factors showed $p>0.05$), and the values of each factor changed only slightly.

Figure 5(c and d) show the results of the MMGS. At time 1, the Cronbach's alpha test yielded a high level of internal consistency for all items of commentary quality (17 items, $N=0.817$, $I=0.876$, $E=0.851$, $IE=0.835$). There was a trend that condition E was rated highest in terms of all factors. Although condition I was rated lowest for all factors. A one-way ANOVA test revealed that there were significant differences between conditions in terms of general usability ($F_{3,176}=19.745$, $p<0.001$), learnability and control ($F_{3,176}=18.344$, $p<0.001$), and quality of interaction with the guide ($F_{3,176}=7.000$, $p<0.001$). Post-hoc tests indicated significant differences between conditions N and E in terms of general usability ($p=0.050$) and quality of interaction with the guide ($p=0.002$). The differences between conditions I and E ($p<0.001$, $p<0.001$, $p<0.001$) and between conditions E and IE ($p<0.001$, $p<0.001$, $p=0.007$) were significant in terms of all factors. At time 2, the Cronbach's alpha test yielded a high level of internal consistency across all items related to the commentary quality (17 items, $N=0.781$, $I=0.871$, $E=0.870$, $IE=0.901$). Condition E was still rated highest overall. A one-way ANOVA test revealed that there were significant differences between conditions in terms of general usability ($F_{3,176}=15.409$, $p<0.001$), learnability and control ($F_{3,176}=13.355$, $p<0.001$), and quality of interaction with the guide ($F_{3,176}=9.734$, $p<0.001$). Post-hoc tests indicated significant differences between conditions N and E in terms of general usability ($p=0.030$) and quality of interaction with the guide ($p<0.001$). The differences between conditions I and E were significant in terms of all factors ($p<0.001$, $p<0.001$, $p<0.001$). Results of paired samples t-tests between times 1 and 2 in all factors of the MMGS are shown in Table 3. There were no significant differences in the values of most factors (3 factors, $N=2/3$, $I=3/3$, $E=2/3$, $IE=2/3$, X/Y: X out of Y factors showed $p>0.05$), and the values of each factor changed only slightly.

5.1.2. Acquisition of audio information and visual information

Figure 6(a and b) show the results of the acquisition of audio information. At time 1, the post-test scores of the

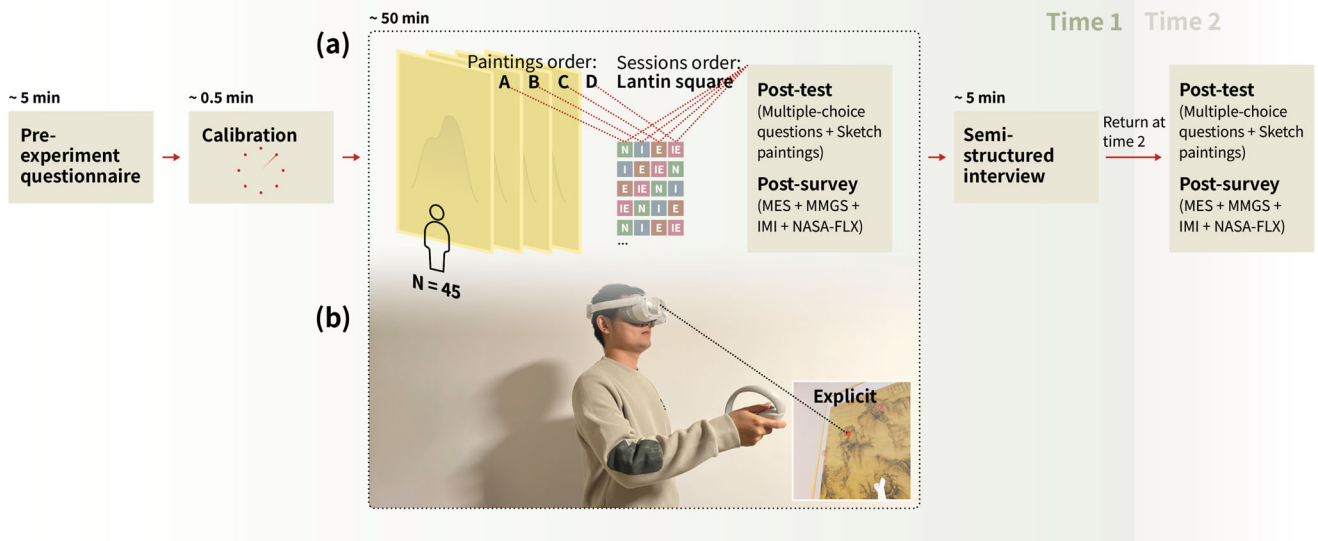


Figure 4. The procedure of the experiment. (a) the use of latin square for the counterbalanced sequence of sessions. N: no eye-tracking; I: implicit; E: explicit; IE: implicit & explicit. The order of the paintings was fixed, and participants viewed only one painting in each virtual museum for each condition. (b) Experimental scenario in condition E.

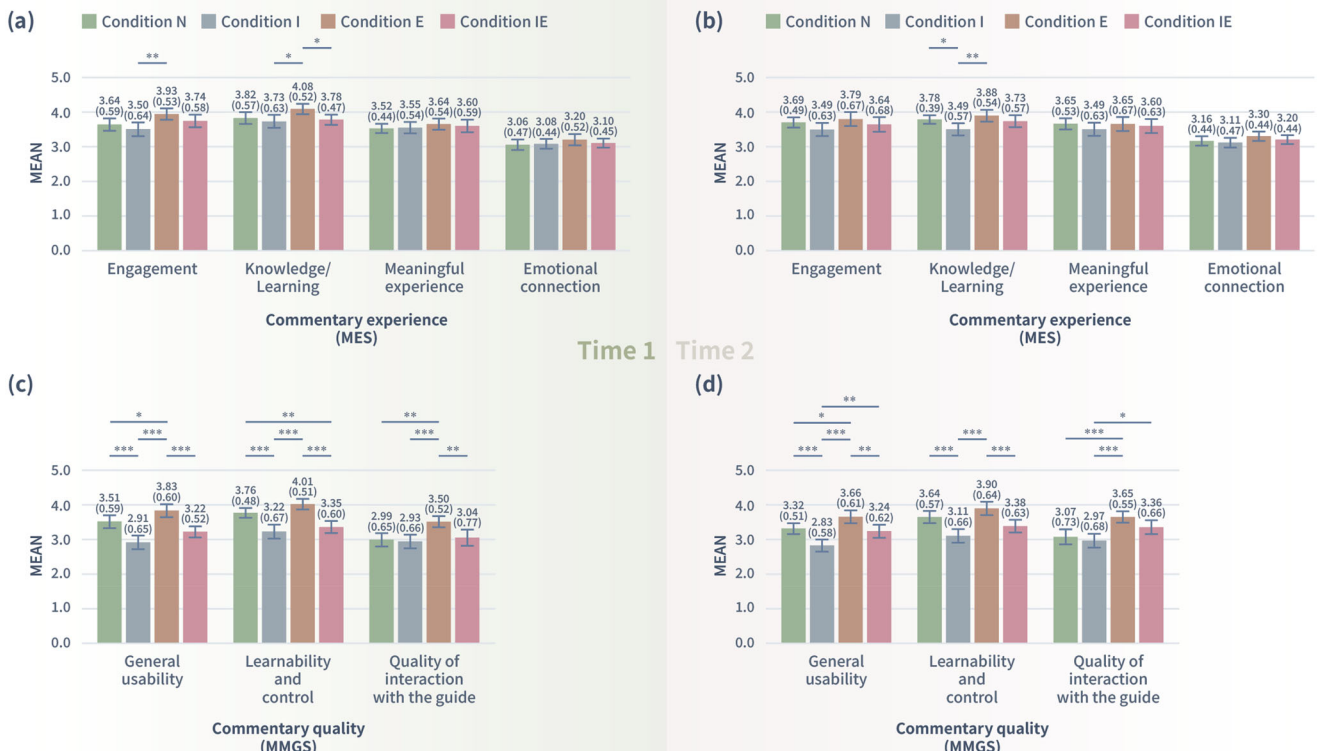


Figure 5. (a) Comparison of MES factors between different conditions at time 1 (with standard deviations). (b) Comparison of MES factors between different conditions at time 2 (with standard deviations). (c) Comparison of MMGS factors between different conditions at time 1 (with standard deviations). (d) Comparison of MMGS factors between different conditions at time 2 (with standard deviations). Error bars mark 95% confidence interval, * indicates significant effect at $p \leq 0.05$, ** indicates significant effect at $p \leq 0.01$, *** indicates significant effect at $p \leq 0.001$.

three methods of gaze-based audio commentaries (conditions I, E, and IE) were lower than the classic audio commentary (condition N). Paired samples t-tests showed that there were significant differences between pre-test and post-test scores in conditions N ($t_{44} = -16.407$, $p < 0.001$), I ($t_{44} = -11.435$, $p < 0.001$), E ($t_{44} = -15.793$, $p < 0.001$), and IE ($t_{44} = -14.012$, $p < 0.001$). A one-way ANOVA test revealed that

there were no significant differences between conditions in the pre-test scores of audio information ($F_{3,176} = 1.725$, $p = 0.163$), but significant differences between conditions in the post-test scores of audio information ($F_{3,176} = 3.937$, $p = 0.009$). Post-hoc tests indicated significant differences between conditions N and I ($p = 0.010$) and between conditions I and E ($p = 0.042$). Further analysis of the post-test scores in the five dimensions

Table 3. Results of paired samples t-tests between times 1 and 2 in all factors of the MES and MMGS.

Condition	Commentary Experience (MES)								Commentary Quality (MMGS)					
	Engagement		Knowledge/Learning		Meaningful Experience		Emotional Connection		General Usability		Learnability and Control		Quality of Interaction with the Guide	
	t_{44}	p	t_{44}	p	t_{44}	p	t_{44}	p	t_{44}	p	t_{44}	p	t_{44}	p
N	-0.688	0.495	0.593	0.556	-1.666	0.103	-1.680	0.100	2.068	0.045*	1.386	0.173	-0.735	0.466
I	0.120	0.905	2.055	0.046*	0.556	0.581	-0.455	0.652	0.788	0.435	1.166	0.250	-0.302	0.764
E	1.549	0.129	2.091	0.042*	-0.049	0.961	-1.533	0.132	2.262	0.029*	1.244	0.220	-1.799	0.079
IE	0.962	0.341	0.514	0.610	0.000	1.000	-1.533	0.132	-0.226	0.823	-0.355	0.724	-2.746	0.009**

Significance coded as such: * $p \leq 0.05$, ** $p \leq 0.01$.

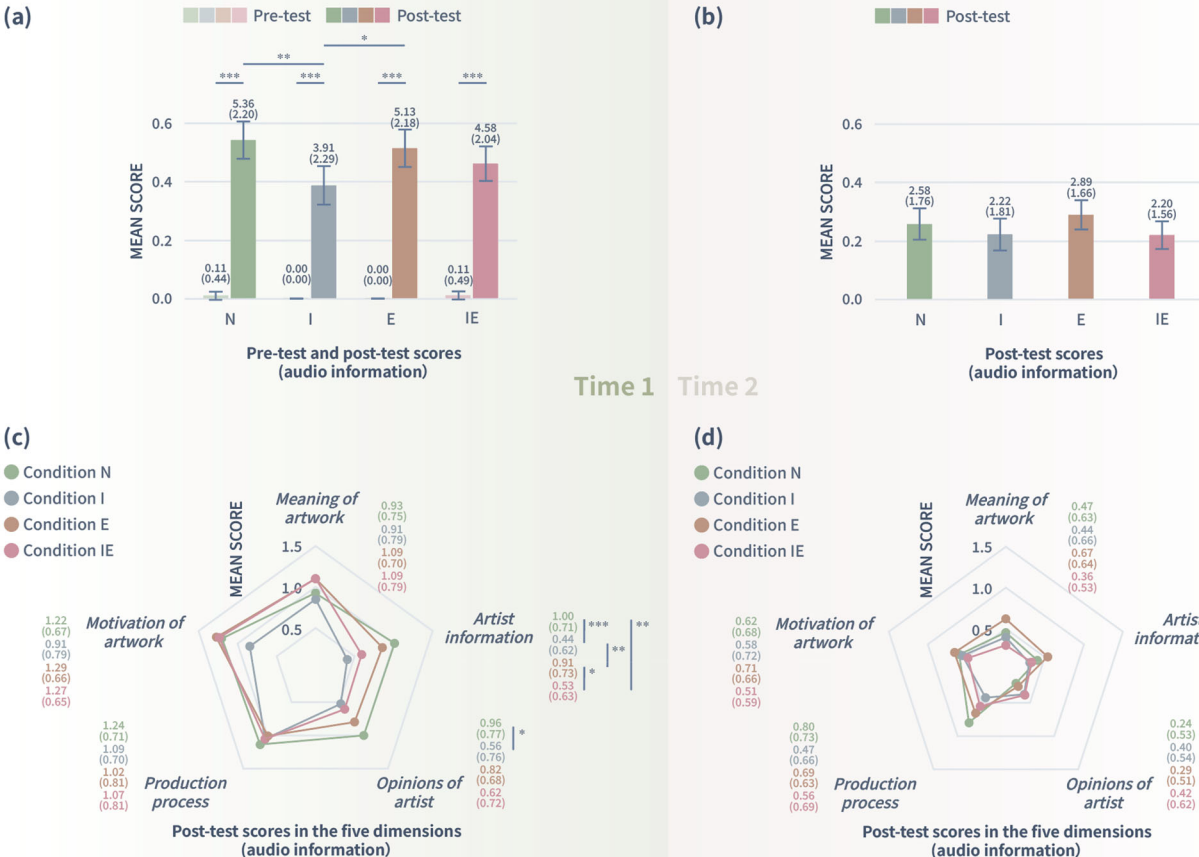


Figure 6. (a) Means of the overall pre-test and post-test scores of audio information between different conditions at time 1 (with standard deviations). (b) Means of the overall post-test scores of audio information between different conditions at time 2 (with standard deviations). (c) Means of the post-test scores of audio information in the five dimensions between different conditions at time 1 (with standard deviations). (d) Means of the post-test scores of audio information in the five dimensions between different conditions at time 2 (with standard deviations). Error bars mark 95% confidence interval, * indicates significant effect at $p \leq 0.05$, ** indicates significant effect at $p \leq 0.01$, *** indicates significant effect at $p \leq 0.001$.

of audio information (Figure 6(c and d)) showed that conditions E and IE were the highest in terms of *meaning of artwork*, condition E was the highest in terms of *motivation of artwork*, and condition N was the highest in terms of *production process*, *opinions of artist*, and *artist information*. A one-way ANOVA test revealed that there were significant differences between conditions in terms of *opinions of artist* ($F_{3, 176} = 2.834, p = 0.040$) and *artist information* ($F_{3, 176} = 7.452, p < 0.001$). Post-hoc tests indicated significant differences between conditions N and I in terms of *opinions of artist* ($p = 0.050$) and *artist information* ($p = 0.001$). The differences between conditions N and IE ($p = 0.007$), between conditions I and E ($p = 0.007$), and between conditions E

and IE ($p = 0.042$) were significant in terms of *artist information*. At time 2, a one-way ANOVA test revealed that there were no significant differences between conditions in the post-test scores of audio information ($F_{3, 176} = 1.669, p = 0.175$). Analysis of the five dimensions of audio information showed that the differences were insignificant between conditions in terms of *meaning of artwork* ($F_{3, 176} = 2.050, p = 0.109$), *motivation of artwork* ($F_{3, 176} = 0.714, p = 0.545$), *production process* ($F_{3, 176} = 2.100, p = 0.102$), *opinions of artwork* ($F_{3, 176} = 1.094, p = 0.353$), and *artist information* ($F_{3, 176} = 1.142, p = 0.334$). Results of paired samples t-tests between times 1 and 2 in all dimensions of the acquisition of audio information are shown in Table 4. There were

Table 4. Results of paired samples t-tests between times 1 and 2 in all dimensions of the acquisition of audio information.

Condition	Post-Test Score in the Five Dimension (Audio Information)											
	Post-Test Score (Audio Information)		Meaning of Artwork		Motivation of Artwork		Production Process		Opinions of Artist		Artist Information	
	t_{44}	p	t_{44}	p	t_{44}	p	t_{44}	p	t_{44}	p	t_{44}	p
N	10.033	0.000***	3.159	0.003**	5.854	0.000***	3.798	0.000***	5.851	0.000***	5.142	0.000***
I	5.048	0.000***	4.311	0.000***	2.406	0.020*	5.371	0.000***	1.416	0.164	0.868	0.390
E	6.143	0.000***	3.617	0.001***	4.950	0.000***	2.909	0.006**	5.154	0.000***	2.945	0.005**
IE	8.988	0.000***	6.080	0.000***	6.107	0.000***	4.069	0.000***	1.706	0.095	1.943	0.058

Significance coded as such: * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$.

significant differences in the post-test scores and in the post-test scores of most dimensions (5 dimensions, $N = 5/5$, $I = 3/5$, $E = 5/5$, $IE = 3/5$, X/Y : X out of Y dimensions showed $p \leq 0.05$), with scores for each dimension decreasing notably.

Figure 7(a and b) show the three examples of sketch paintings and Figure 7(c and e) show the results of the acquisition of visual information. At time 1, the post-test scores of the three methods of gaze-based audio commentaries (conditions I, E, and IE) were higher than the classic audio commentary (condition N). A one-way ANOVA test revealed that there were significant differences between conditions in the post-test scores of visual information ($F_{3, 176} = 3.988$, $p = 0.009$). Post-hoc tests indicated significant differences between conditions N and E ($p = 0.009$) and between conditions I and E ($p = 0.038$). Analysis of the percentage of audio-related scores (Figure 7(d and f)) showed that the three methods of gaze-based audio commentaries (conditions I, E, and IE) were higher than the classic audio commentary (condition N). A one-way ANOVA test revealed that there were significant differences between conditions ($F_{3, 176} = 5.256$, $p = 0.002$). Post-hoc tests indicated significant differences between conditions N and I ($p = 0.033$) and between conditions N and E ($p = 0.005$). Further analysis of the post-test scores in the three dimensions of visual information (Figure 7(g and h)) showed that condition E was the highest in terms of *position* and *dimension*, and condition I was the highest in terms of *detail*. A one-way ANOVA test revealed that there were significant differences between conditions in terms of *position* ($F_{3, 176} = 6.997$, $p < 0.001$), *dimension* ($F_{3, 176} = 5.259$, $p = 0.002$), and *detail* ($F_{3, 176} = 4.166$, $p = 0.007$). Post-hoc tests indicated significant differences between conditions N and E ($p = 0.001$, $p = 0.002$) and between conditions I and E ($p < 0.001$, $p = 0.008$) in terms of *position* and *dimension*. The differences between conditions I and E ($p = 0.044$) and between conditions I and IE ($p = 0.006$) were significant in terms of *detail*. At time 2, it was notable that the results of sketch paintings varied widely between participants. Some participants forgot the visual elements of one or more paintings altogether ($N = 11$), while some participants sketched similarly to their performance in time 1 ($N = 8$). It was also noteworthy that a number of participants demonstrated a tendency to associate the visual elements with paintings that were not the correct ones ($N = 10$). For participants who performed similarly in time 2 compared to time 1, we also inquired whether they had engaged in content-related daily experiences after time 1. These daily experiences included but were not limited to recalling paintings during free time,

discussing cultural heritages with peers, and appreciating natural scenery. More than 50% of them reported that they had conducted these daily experiences ($N = 5$). Additionally, a one-way ANOVA test revealed that there were no significant differences between conditions in the post-test scores of visual information ($F_{3, 176} = 0.088$, $p = 0.966$), the percentage of audio-related scores ($F_{3, 176} = 0.411$, $p = 0.745$), and the three dimensions of visual information: *position* ($F_{3, 176} = 0.122$, $p = 0.947$), *dimension* ($F_{3, 176} = 0.070$, $p = 0.976$), and *detail* ($F_{3, 176} = 0.577$, $p = 0.631$). Results of paired samples t-tests between times 1 and 2 in all dimensions of the acquisition of visual information are shown in Table 5. There were significant differences in the post-test scores and in the post-test scores of all dimensions (3 dimensions, $N = 3/3$, $I = 3/3$, $E = 3/3$, $IE = 3/3$, X/Y : X out of Y dimensions showed $p \leq 0.05$), with scores for each dimension decreasing notably.

5.1.3. Intrinsic motivation and cognitive load

Figure 8(a and b) show the results of the IMI. At time 1, the Cronbach's alpha test yielded a high level of internal consistency for all items of intrinsic motivation (22 items, $N = 0.736$, $I = 0.728$, $E = 0.720$, $IE = 0.764$). There was a trend that condition E was rated highest in terms of interest/enjoyment, perceived competence, and perceived choice. In terms of pressure/tension, condition N was rated lowest. A one-way ANOVA test revealed that there were significant differences between conditions in terms of perceived competence ($F_{3, 176} = 5.237$, $p = 0.002$) and pressure/tension ($F_{3, 176} = 3.226$, $p = 0.024$). Post-hoc tests indicated significant differences between conditions N and I in terms of perceived competence ($p = 0.038$) and pressure/tension ($p = 0.036$). The differences between conditions I and E were significant in terms of perceived competence ($p = 0.002$). At time 2, the Cronbach's alpha test yielded a high level of internal consistency across all items related to the intrinsic motivation (22 items, $N = 0.809$, $I = 0.855$, $E = 0.816$, $IE = 0.824$). A one-way ANOVA test revealed that there were significant differences between conditions in terms of perceived competence ($F_{3, 176} = 3.951$, $p = 0.009$). Post-hoc tests indicated significant differences between conditions I and E in terms of perceived competence ($p = 0.005$). Results of paired samples t-tests between times 1 and 2 in all factors of the IMI are shown in Table 6. There were no significant differences in the values of most factors (4 factors, $N = 2/4$, $I = 3/4$, $E = 2/4$, $IE = 3/4$, X/Y : X out of Y factors showed $p > 0.05$), and the values of each factor changed only slightly.

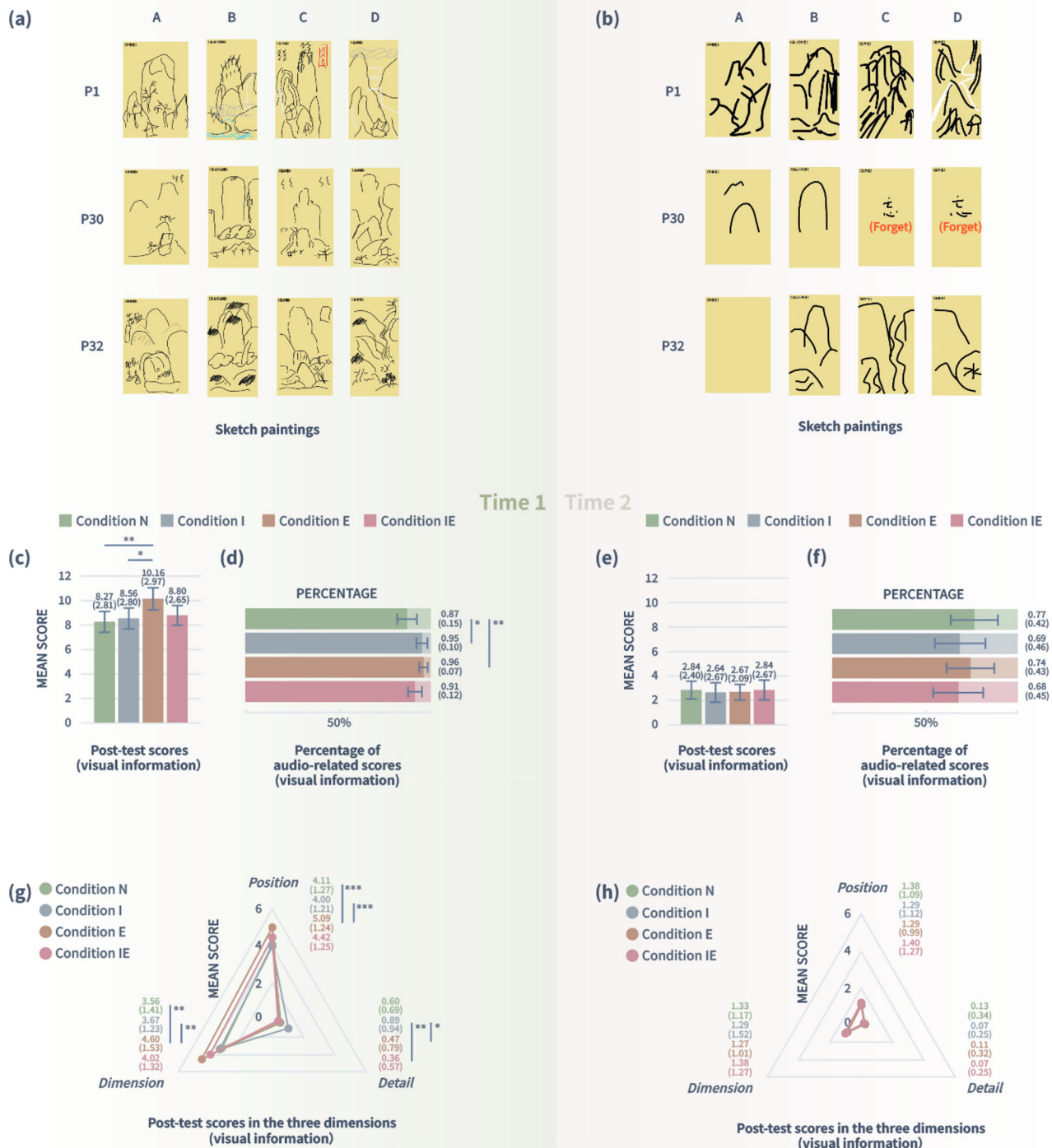


Figure 7. (a) The three examples of sketch paintings between different conditions at time 1. (b) The three examples of sketch paintings between different conditions at time 2. (c) Means of the overall post-test scores of visual information between different conditions at time 1 (with standard deviations). (d) Means of the percentage of audio-related scores between different conditions at time 1 (with standard deviations). (e) Means of the overall post-test scores of visual information between different conditions at time 2 (with standard deviations). (f) Means of the percentage of audio-related scores between different conditions at time 2 (with standard deviations). (g) Means of the post-test scores in the three dimensions of visual information between different conditions at time 1 (with standard deviations). (h) Means of the post-test scores in the three dimensions of visual information between different conditions at time 2 (with standard deviations). Error bars mark 95% confidence interval, * indicates significant effect at $p \leq 0.05$, ** indicates significant effect at $p \leq 0.01$, *** indicates significant effect at $p \leq 0.001$.

Figure 8(c and d) show the results of the NASA-FLX. At time 1, the Cronbach's alpha test yielded a high level of internal consistency for all items of cognitive load (6 items, $N = 0.839$, $I = 0.684$, $E = 0.701$, $IE = 0.718$). There was a trend that the three methods of gaze-based audio commentaries (conditions I, E, and IE) were rated higher than the

classic audio commentary (condition N). A one-way ANOVA test revealed that there were significant differences between conditions ($F_{3,176} = 7.934$, $p < 0.001$). Post-hoc tests indicated significant differences between conditions N and I ($p = 0.001$), between conditions I and E ($p = 0.003$), between conditions E and IE ($p = 0.027$), and between conditions N

Table 5. Results of paired samples t-tests between times 1 and 2 in all dimensions of the acquisition of visual information.

Condition	Post-Test Score (Visual Information)		Percentage of Audio-Related Score		Post-Test Score in the Three Dimension (Visual Information)					
					Position		Dimension		Detail	
	t_{44}	p	t_{44}	p	t_{44}	p	t_{44}	p	t_{44}	p
N	12.110	0.000***	1.670	0.102	12.362	0.000***	10.104	0.000***	5.326	0.000***
I	12.488	0.000***	3.875	0.000***	13.907	0.000***	9.331	0.000***	6.222	0.000***
E	15.004	0.000***	3.427	0.001***	18.107	0.000***	13.110	0.006**	3.084	0.004**
IE	11.690	0.000***	3.367	0.002**	12.127	0.000***	10.640	0.000***	3.532	0.001***

Significance coded as such: ** $p \leq 0.01$, *** $p \leq 0.001$.

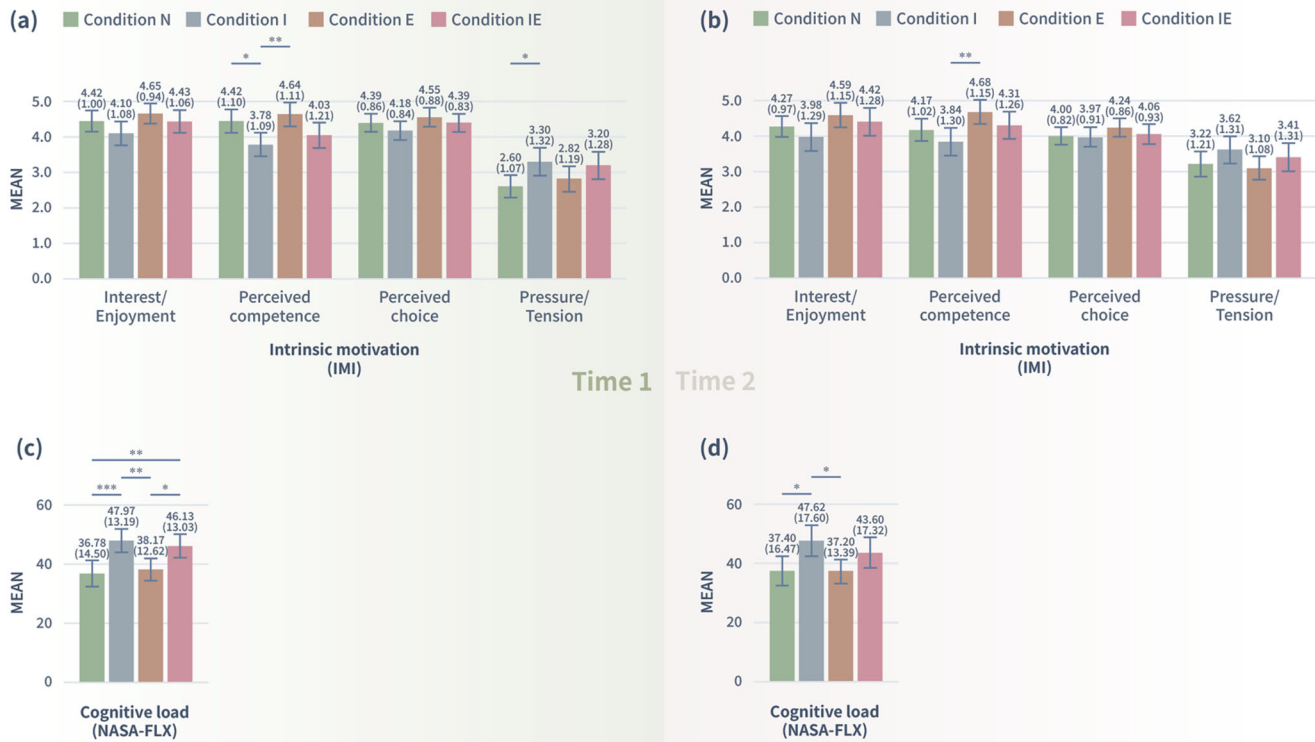


Figure 8. (a) Comparison of IMI factors between different conditions at time 1 (with standard deviations). (b) Comparison of IMI factors between different conditions at time 2 (with standard deviations). (c) Average NASA-TLX values between different conditions at time 1 (with standard deviations). (d) Average NASA-TLX values between different conditions at time 2 (with standard deviations). Error bars mark 95% confidence interval, * indicates significant effect at $p \leq 0.05$, ** indicates significant effect at $p \leq 0.01$, *** indicates significant effect at $p \leq 0.001$.

and IE ($p = 0.006$). At time 2, the Cronbach's alpha test yielded a high level of internal consistency across all items related to the cognitive load (6 items, $N = 0.811$, $I = 0.853$, $E = 0.767$, $IE = 0.880$). A one-way ANOVA test revealed that there were significant differences between conditions ($F_{3,176} = 4.368$, $p = 0.005$). Post-hoc tests indicated significant differences between conditions N and I ($p = 0.017$) and between conditions I and E ($p = 0.014$). Results of paired samples t-tests between times 1 and 2 in the NASA-FLX are shown in Table 6. There were no significant differences in the values of cognitive load, and the values of each condition changed only slightly.

5.2. Qualitative results

We conducted theme-based content analysis on the interview transcription data, focusing on aspects of experience and learning.

In terms of experience, although the classic audio commentary was considered uninteresting due to its lack of interaction, some participants preferred it for its easy and smooth experience, e.g., "Condition N is easy to control because I don't need eyes to maneuver it and it goes straight to voice announcements" (P22), "Condition N doesn't affect my train of thought. I can listen and view at the same time or listen and then view" (P27). Gaze-based audio commentaries engaged participants by providing a sense of participation through eye-controlled interaction. As P2 said, "Gaze-based audio commentaries are a method of engaging me and enabling me to learn about the painting on my own initiative." Among the gaze-based audio commentaries, condition I was perceived by participants as lack of control, which imposed a burden on them. As P33 stated, "Since I need to think about how to trigger the voice, which causes me a lot of burden." In contrast to condition I, the speaker icons in condition E made the operation intuitive for participants, leading them to regard this method as easier. As P19 mentioned, "I feel that condition E is more

Table 6. Results of paired samples t-tests between times 1 and 2 in all factors of the IMI and NASA-FLX.

Condition	Intrinsic Motivation (IMI)									
	Interest/Enjoyment		Perceived Competence		Perceived Choice		Pressure/Tension		Cognitive Load (NASA-FLX)	
	t_{44}	p	t_{44}	p	t_{44}	p	t_{44}	p	t_{44}	p
N	1.229	0.225	1.755	0.086	3.297	0.002**	-3.907	0.000***	-0.311	0.758
I	0.647	0.521	-0.343	0.733	1.400	0.169	-2.215	0.032*	0.185	0.854
E	0.348	0.730	-0.320	0.750	2.764	0.008**	-2.416	0.020*	0.655	0.516
IE	0.056	0.956	-1.688	0.098	2.404	0.020*	-1.536	0.132	1.162	0.251

Significance coded as such: * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$.

straightforward, I don't have to think about where exactly I am going to look, it just tells you already." However, participants also felt that the speaker icons disrupted their vision, hindering them from fully appreciating the painting. As P22 announced, "Since condition E includes visible red speaker icons, it may intrude on the initial viewing experience of the painting." Furthermore, condition IE integrated elements of both conditions I and E, which participants perceived as a way to offset the respective limitations of these individual methods. Specifically, participants believed that it improved the lack of control in condition I, while simultaneously avoiding disrupted vision in condition E, e.g., "I find that condition IE seemed to make the system more responsive to my needs" (P45), "Condition IE then shows the speaker icon after I trigger it, which doesn't disrupt my vision of the whole painting" (P25).

In terms of learning, participants generally perceived the classic audio commentary as presenting all the information at once, leaving little time for them to react and contemplate the content. As P7 said, "Condition N's audio commentaries broadcast at once, which can make me feel a bit rushed to receive the information." In addition, the lack of a clear association between the classic audio commentary and the corresponding areas of the painting made it difficult for participants to connect the audio commentaries with the visual elements. As P25 stated, "Sometimes you might not know what part of the painting it's talking about." In contrast to the classic audio commentary, gaze-based audio commentaries segmented the audio commentaries to correspond with visual elements individually, which participants perceived as a more effective approach to learning. As P10 mentioned, "Gaze-based audio commentaries are more targeted. I can focus directly on the content the voice is describing, which enhances my learning." Among the gaze-based audio commentaries, participants considered condition I to be more suitable for free exploration, as it did not indicate the locations of the audio commentaries. As P7 announced, "Condition I allows me the freedom to explore, and I can appreciate the painting from my own perspective." However, this might lead to participants missing information by not triggering all the audio commentaries. As P40 argued, "I don't know exactly how many audio commentaries are on it, I could miss a lot." The speaker icons above the painting in condition E were perceived by participants as indirect cues to the focus and composition of the painting. As P29 said, "Condition E has pronounced speaker icons, which can give an early indication of the focus of the painting." Additionally, participants generally felt that condition IE combined the

advantages of conditions I and E, allowing them the freedom to explore while receiving prompts that could trigger audio commentaries, which supported their self-directed learning, e.g., "When I'm more interested in an area, the area alerts me to the possibility of an audio commentary" (P28), "Condition IE allows me to decide whether or not to listen based on what interests me" (P14).

During the interviews, participants proposed several interesting and inspiring design solutions to address the challenges associated with gaze-based audio commentaries.

1. For condition I: provide distance information to guide gaze. Specifically, participants suggested that audio commentaries should not simply add the spatial positioning cue of the visual element, but should provide information about the distance between the gaze point and the visual element to guide the gaze to the visual element.
2. For condition E: adjust the display and hiding of the speaker icons (static buttons) based on the visitor's distance from the painting. Specifically, participants recommended that speaker icons should remain hidden when the visitor viewed the composition of the painting from a distance and appear only when he moved closer to focus on details.
3. For condition IE: add breathing speaker icons (static buttons). Specifically, participants proposed that speaker icons should not only appear through gazing but also intermittently disappear and reappear. This dynamic behavior could serve as a subtle clue for visitors, helping them identify where audio commentaries were available.
4. For gaze-based audio commentaries: switch from explicit to implicit method. Specifically, based on participants' feedback, the audio commentary system would begin in the explicit method, where each speaker icon, once activated, disappeared permanently. After a speaker icon disappeared, its associated AOI switched to the implicit method. Once all speaker icons had been activated and disappeared, the entire system transitioned fully into the implicit method.

6. Discussion

In this section, we discuss the results in relation to the three research questions of this study and provide design insights and strategies from an integrated perspective. Additionally, the current limitations and directions for future research are discussed.

6.1. Discussion of results

RQ1: Do the three methods of gaze-based audio commentaries enhance the experience and quality of audio commentaries?

At time 1, the results of MES revealed that condition E could provide users with a more positive commentary experience than condition I in terms of engagement and knowledge/learning, and than condition IE in terms of knowledge/learning. This showed that the explicit method enhanced users' engagement and learning experience. These results were expected, as participants reported during the interviews that they felt more engaged ($N=2$) and more effective ($N=8$) with condition E. We believed that it could be the intuitive interactive nature of the explicit method, which enabled condition E to offer a better commentary experience. However, conditions I and IE were unsatisfactory in terms of all factors of commentary experience due to their lack of intuitiveness. During the interviews, participants reported that the lack of intuitiveness caused uncertainty about their gaze directions, making them feel a lack of control and requiring considerable time to familiarize themselves with the system, which negatively impacted their experiences ($N=26$). At time 2, the results suggested that users' perception of the differences between conditions in commentary experience in terms of knowledge/learning, meaningful experience, and emotional connection largely remained the same after one month. Between times 1 and 2, the results indicated almost no significant change in users' perception of all factors of commentary experience after one month. Overall, the users' perception of the commentary experience was largely long-term. This further highlighted the value of the explicit method in improving users' engagement and learning experience.

At time 1, the results of MMGS revealed that condition E was considered the best in terms of general usability, learnability and control, and quality of interaction with the guide. This showed that the explicit method improved audio commentaries' usability, learnability, and interactivity. These results were expected, as participants reported during the interviews that they felt condition E was easier to use ($N=6$), more controlled ($N=13$), and clearer ($N=23$). Obviously, this was due to the intuitiveness of the explicit method. However, the commentary quality of conditions I and IE remained unsatisfactory. This might be due to the lack of intuitiveness and the negative commentary experience, leading participants to perceive them as of low commentary quality. At time 2, the results suggested that users' perception of the differences between conditions in all factors of commentary quality largely remained the same after one month. Between times 1 and 2, the results indicated almost no significant change in users' perception of all factors of commentary quality after one month. Overall, the users' perception of the commentary quality was largely long-term. This further highlighted the value of the explicit method in improving audio commentaries' usability, learnability, and interactivity.

RQ2: Do the three methods of gaze-based audio commentaries improve users' acquisition of audio information and visual information?

In terms of the acquisition of audio information, at time 1, the results revealed that gaze-based audio commentaries had a negative impact on users' acquisition of audio information compared to classic audio commentaries. This finding is not aligned with the widely acknowledged view that interaction contributes to users' retention of information (Sugiura et al., 2022). One possible explanation for this difference was that the act of eye-controlled interaction might divert some of the user's attentional resources, which could impede their formation of conscious memories (Chun & Turk-Browne, 2007). This might also be due to the user's limited familiarity with eye-controlled interaction in the pre-experimental research, as familiarity with technologies could help individuals access information more efficiently (C.-C. Chen et al., 2024). Among the gaze-based audio commentaries, participants reported during the interviews that condition I was lack of control ($N=11$), while condition IE experienced similar issues but to a lesser extent than condition I ($N=9$). This further diminished users' ability to familiarize themselves with and master this kind of interaction, leading to the learning outcomes of conditions I and IE being less effective than those of condition E. Additionally, the results of the five dimensions of audio information demonstrated that gaze-based audio commentaries limited users' learning of *production process*, *opinions of artist*, and *artist information*. This might be because audio content, such as *opinions of artist* and *artist information*, was challenging to intuitively connect with the visual elements in paintings, making it redundant information. The redundant information disrupted the congruence between visual and auditory stimuli, hindering the additive learning effect created by the combination of images and verbal (Paivio & Csapo, 1969, 1973; Thompson & Paivio, 1994). Therefore, it is advisable to select audio content that can be associated with visual elements for gaze-based audio commentaries, such as *meaning of artwork* and *motivation of artwork*. At time 2, the results suggested that after one month, there were no significant differences between conditions in the user's retention of all dimensions of audio information. Between times 1 and 2, the results indicated an almost significant reduction in users' retention of all dimensions of audio information after one month. Overall, gaze-based audio commentaries largely did not support users in forming lasting memories of audio information. This indicated that gaze-based audio commentaries lacked long-term value for users' acquisition of audio information. Part of the reason might be that the participants in this study lacked prior knowledge of and interest in traditional Chinese landscape paintings, which were essential prerequisites for forming lasting memories. Because, on the one hand, learning in museums is built on the consolidation and transformation of prior knowledge; on the other hand, a lack of interest may lead most visitors to disengage from content-related activities after leaving the museum (Falk & Dierking, 2000).

In terms of the acquisition of visual information, at time 1, the results revealed that users could achieve better acquisition of visual information by using gaze-based audio commentaries, particularly the explicit method, compared to

classic audio commentaries. This might be attributed to the way eye-controlled interaction guided the line of sight to the visual element associated with the audio commentary, optimizing the allocation of the users' cognitive resources and enhancing multisensory learning, thereby allowing them to remember more visual information (Kim et al., 2008; Mayer, 2005; Shams & Seitz, 2008). Participants also reported during the interviews that gaze-based audio commentaries with targeted guidance on visual elements helped them remember better ($N=11$). The results of the percentage of audio-related scores indicated that gaze-based audio commentaries, especially conditions I and E, enabled users to better memorize the visual elements associated with the audio commentaries. However, it also resulted in users remembering fewer visual elements that were unrelated to the audio commentaries, potentially missing out on aspects of the painting that might have truly interested them. Participants also expressed concerns during the interviews that gaze-based audio commentaries might restrict their freedom to explore visual elements of personal interest ($N=5$). These findings align with previous studies suggesting that audio guides (also audio commentaries) could impede independent thought (Bauer-Krösbacher, 2013), and further demonstrate that gaze-based audio commentaries might exacerbate this issue. Additionally, the results of the three dimensions of visual information revealed that condition I facilitated the learning of *detail*, and condition E helped to enhance the learning of *position* and *dimension*. Participants also mentioned during the interviews that condition I facilitated careful observation of the painting ($N=2$), while condition E provided tips regarding the composition of the painting ($N=3$). These demonstrated that conditions I and E each had advantages for the acquisition of visual information among the gaze-based audio commentaries. At time 2, the results suggested that after one month, there were no significant differences between conditions in the user's retention of all dimensions of visual information. However, the gap in performance on sketch paintings among participants widened. Interviews suggested that some participants gradually forgot or confused the visual information after the experiment, while others reinforced what they had learned in daily experiences. These findings align with the assertion of Falk and Dierking (2000) that memory and learning are dynamic processes shaped by daily experiences. Between times 1 and 2, the results indicated a significant reduction in users' retention of all dimensions of visual information after one month. Overall, gaze-based audio commentaries did not support users in forming lasting memories of visual information. This indicated that gaze-based audio commentaries lacked long-term value for facilitating users' acquisition of visual information. Part of the reason might be the same as the previous explanation for the acquisition of audio information.

RQ3: *Do the three methods of gaze-based audio commentaries enhance intrinsic motivation and reduce cognitive load?*

In terms of intrinsic motivation, at time 1, the results revealed that condition E had a more positive performance on intrinsic motivation in terms of perceived competence compared to condition I. This might be because condition E provided users with a better commentary experience and

commentary quality. In addition, we believed that this was a contributing factor to the positive outcomes of the acquisition of audio information and visual information in condition E compared to conditions I and IE, as learning might be more robust when there was higher intrinsic motivation (Falk & Dierking, 1997). Additionally, as expected, conditions I and E were unsatisfactory in terms of all factors due to the poor commentary experience and commentary quality caused by the lack of intuitiveness. At time 2, the results suggested that users' perception of the differences between conditions in intrinsic motivation in terms of interest/enjoyment, perceived competence, and perceived choice largely remained the same after one month. Between times 1 and 2, the results indicated almost no significant change in users' perception of all factors of intrinsic motivation after one month. Overall, the users' perception of the intrinsic motivation was largely long-term. This further highlighted the value of the explicit method in improving intrinsic motivation.

In terms of cognitive load, at time 1, the results revealed that gaze-based audio commentaries brought higher cognitive load compared to classic audio commentaries. We suspected that this was because eye-controlled interaction could limit users' ability to allocate necessary attentional resources to the audio commentaries (Krukar & Dalton, 2020), leading to an increased cognitive load for users. Among the gaze-based audio commentaries, condition E imposed the lowest cognitive load, while condition I imposed the highest cognitive load. This might be attributed to the differences in the intuitiveness of interaction between the implicit method and the explicit method, as participants reported during the interviews that condition E was easy to use because of its intuitiveness ($N=24$), while condition I was difficult to manipulate due to its lack of intuitiveness ($N=13$). We also observed in the experiment that, although participants were not required to actively control the audio commentaries in the implicit method, they instinctively tried to search for and trigger the audio commentaries, which caused them distress. This might be the underlying reason for the increased cognitive load in condition I. This finding is inconsistent with the expectation of interface-free interaction, which argues that applying sensors to enable the computer to implicitly access information reduces the burden on the user (Krishna, 2015). Additionally, condition IE imposed a relatively high cognitive load compared to conditions N and E. The most frequently mentioned reason by participants during the interviews was the higher complexity of the implicit & explicit method compared to the other methods ($N=7$), which resulted in a greater operational burden. This was also the reason why condition IE performed poorly in terms of experience and learning. Overall, care should be taken when applying gaze-based audio commentaries, particularly the implicit method, due to their potential negative impact on cognitive load. At time 2, the results suggested that users' perception of the differences between conditions in cognitive load largely remained the same after one month. Between times 1 and 2, the results indicated no significant change in users' perception of the cognitive load after one month. Overall, the users' perception of the cognitive load was

largely long-term. This further highlighted the risks associated with gaze-based audio commentaries in increasing cognitive load.

6.1.1. Design insights and strategies of an integrated perspective

The results revealed that gaze-based audio commentaries might impede the acquisition of audio information, limit free viewing, and increase cognitive load. Among the gaze-based audio commentaries, the performance of conditions I and IE was generally unsatisfactory in most factors compared to condition E. Therefore, it is crucial to establish clear strategies to guide the design of gaze-based audio commentaries, ensuring negative effects are minimized and value maximized. Synthesizing the above discussions, we propose the following insights and strategies for the design of gaze-based audio commentaries.

1. The implicit method presents several challenges, including insufficient intuitiveness, lack of control, and hard to use. These challenges are partly due to the inherent disadvantage of the implicit method being invisible. Therefore, we suggest integrating the implicit and explicit methods as a solution to improve this method. In addition, the challenges encountered may also stem from the simplistic design of the implicit method used in this study, which lacks gaze guidance. Based on participants' feedback, it is hypothesized that incorporating more detailed gaze guidance in audio commentaries, such as information about the distance between the gaze point and the visual element, could effectively direct the user's gaze.
2. Although the explicit method is the most ideal among the gaze-based audio commentaries, it has a notable drawback: speaker icons as static buttons disrupt the vision. While this drawback cannot be entirely resolved, there are several ways to improve it. First, reducing the number of speaker icons can minimize the occlusion areas, which could involve merging audio commentaries or retaining only the more important audio commentaries. Second, speaker icons could be made more transparent or designed in simpler shapes, such as spheres. Last, instead of displaying speaker icons above the painting continuously, they could be hidden when the visitor views the painting from a distance.
3. The design of the implicit & explicit method in this study involves an initial implicit observations phase, followed by direct control, representing a straightforward integration of implicit and explicit methods. Unexpectedly, while the implicit & explicit method was intended to combine the advantages of both methods, it did not receive as positive evaluations as the explicit method overall. Its evaluation mostly fell between the implicit and explicit methods, as it offered a greater sense of control compared to the implicit method but was more complex than the explicit method. Participants suggested potential improvements to make the implicit & explicit method more effective, see (3) and (4) in the qualitative results above. Based on these, we suggest that the implicit & explicit method necessitates a targeted interaction framework to guide the design process, warranting further research.
4. Gaze-based audio commentaries have the advantage of enhancing the acquisition of visual information, and the implicit and explicit methods among them are suitable for different learning objectives. The implicit method is more suitable for learning the details of visual information, while the explicit method is more appropriate for learning the position and dimension of visual information. Both implicit and explicit methods can effectively guide users in the acquisition of visual information linked to audio commentaries. In future designs, these insights can be employed to select appropriate methods of gaze-based audio commentaries based on specific learning content and objectives. Moreover, gaze-based audio commentaries' advantage indicates their potential applicability in educational settings (Okolo et al., 2011; Sederberg, 2013). For example, in the painting classroom setting, gaze-based audio commentaries can be integrated with VR HMDs to serve as teaching tools for students. This method not only brings museum resources into classrooms (Gano & Kinzler, 2011) but also enhances the learning of visual content from paintings using eye-controlled interactions.
5. In order to implement a differentiated strategy for gaze-based audio commentaries, it is necessary to differentiate between the different identities of visitors (Falk, 2009). We can conclude from the findings that explorers might be better suited to use the implicit and implicit & explicit methods, and experience seekers, rechargers, and professional/hobbyists might be better suited to use the explicit method. In addition, to cater to a wider range of visitors, museums could offer a variety of methods of gaze-based audio commentaries, allowing visitors to choose the one that best suits their preferences.
6. For the acquisition of audio information, we suggest the use of narratives to construct engaging stories for audio commentaries, as participants suggested during the interviews that the audio information would be more effectively remembered if it was presented in a storytelling format ($N=3$). Some previous studies also suggested that storytelling could arouse curiosity and positively affect memorability (Kang et al., 2008). For the acquisition of visual information, we suggest enriching the content of audio commentaries with audio descriptions and soundscapes, as these methods are believed to help people associate and recall details of images (Hutchinson & Eardley, 2021, 2024). Additionally, it is also possible to further increase the depth of the content in audio commentaries, especially by setting up different depths of content for visitors with different knowledge bases and learning abilities, which is considered to be one of the means to improve the learning experience and outcome in museums (Falk & Dierking, 2000).
7. During the interviews, participants also offered suggestions for enhancing the functionality of gaze-based

audio commentaries, such as the ability to zoom in and out on paintings, the option to pause and resume audio commentaries, and the capacity to answer questions and communicate. These suggestions involve integrating eye-controlled interaction with other modalities to achieve multimodal interaction, such as combining gaze with gestures or voice. On the one hand, this approach enables visitors to use multiple natural methods of interaction, which is considered to facilitate an immersive, dynamic, and edutainment museum experience (Dondi & Porta, 2023; Raptis et al., 2021). On the other hand, it allows visitors to access information through multiple methods of interaction, aligning with Falk and Dierking's (2000) recommendation to provide diverse entry points for accessing information, which supports museum learning. Therefore, we consider this a promising strategy for improving museum experience and learning, warranting further design practice and experimental evaluation.

6.2. Current limitations and future work

While our study has limitations that may pose potential threats to its validity (Wohlin et al., 2012), they can also offer valuable insights for future research aimed at gaze-based audio commentaries. First, the interconnectedness of visual elements (AOIs) in a painting may influence visitors' gaze patterns (Quian Quiroga & Pedreira, 2011; Villani et al., 2015). Building contextual relationships between these AOIs could further enhance visitors' understanding of the painting's background (Raptis et al., 2021). Therefore, future research should explore these potential relationships in depth and incorporate them into the design of gaze-based audio commentaries, providing corresponding gaze guidance. Second, although gaze-based audio commentaries offer visitors a near-intelligent experience, they are not yet truly intelligent. Further research could integrate artificial intelligence technologies with gaze data to enhance the context awareness of gaze-based audio commentaries, making them more adaptive to the visitor's gaze behavior and motivation. Third, it is common for classic audio commentaries in physical museums to coexist with social interactions. Although the gaze-based audio commentaries in this study were designed for individual use, they could potentially enhance the social experience in museums. For instance, visitors might hear the same audio commentary when their gazes are synchronized on a specific area of a painting. When they look away from the painting to engage with one another, the audio commentary pauses, avoiding any impact on their conversation. Further research could explore this phenomenon in more detail and assess the impact of gaze-based audio commentaries on the social experience within museums. Finally, this study did not account for differences in visitors regarding motivation, expectation, interest, and prior knowledge, among the most significant factors influencing visitors' experience and learning (Falk & Dierking, 2000). Furthermore, the participant group in this study was not diverse, consisting primarily of college students, which might

limit the ecological validity (Winkel, 1985) of the results. In future research, it will be necessary to include a broader range of participants, such as museum professionals and visitors with varying levels of art knowledge, to further explore corresponding gaze-based audio commentary solutions.

7. Conclusion

To our knowledge, this is the first study that aims to explore the potential of eye-controlled interaction in audio commentaries from a paradigm and evaluation perspective. We proposed three methods of gaze-based audio commentaries based on implicit observations and direct control from the eye-controlled interaction paradigms and designed the corresponding prototypes, including implicit, explicit, and implicit & explicit. They were evaluated with a baseline condition of no eye-tracking through a controlled study in virtual museums exhibiting traditional Chinese landscape paintings. The evaluation focused on commentary experience, commentary quality, audio information learning, visual information learning, intrinsic motivation, and cognitive load. Results revealed that gaze-based audio commentaries effectively facilitated users' acquisition of visual information but somewhat hindered the acquisition of audio information, restricted free viewing, and increased cognitive load. Among the gaze-based audio commentaries, the explicit method was found to be better in terms of commentary experience, commentary quality, and intrinsic motivation, which were largely long-term. In terms of the commentary experience, the explicit method enhanced users' engagement and learning experience. In terms of the commentary quality, the explicit method improved audio commentaries' usability, learnability, and interactivity. However, it had the notable drawback of disrupting the vision. The implicit method supported users in learning the details of visual information, while the explicit method aided in learning the position and dimension of visual information. However, both of them lacked long-term value for users' retention of the information. The implicit & explicit method aimed to address the limitations of both individual methods, but the effects were still less optimal compared to the explicit method alone, due to its higher complexity. Finally, we provided design insights and strategies for enhancing the design of the three methods of gaze-based audio commentaries, such as incorporating more detailed guidance, adjusting the display of static buttons, and altering the integration of implicit and explicit methods.

Although this study proposes gaze-based audio commentaries for virtual museums using VR technology, they can also be applied to physical museum environments or other open educational spaces using AR/MR technology to maximize their value. It is important to note that this study's design insights and strategies primarily target VR environments. Although the virtual museums in this study were designed to closely replicate physical museums, certain variables were introduced that may not exist in physical settings. For instance, navigate using a controller in virtual museums, unlike the unrestricted movement available in physical

museums. In addition, the resolution and realism of paintings in virtual museums do not match the quality of the originals displayed in physical museums. Therefore, careful consideration is needed when generalizing these design insights and strategies from VR to physical settings. However, based on the limitations of virtual museums in this study, visitors may have more flexible navigation patterns (Clemenson et al., 2020; Hejtmanek et al., 2020) and more positive emotions and engagement levels (Baradaran Rahimi et al., 2022; Kennedy et al., 2021) in physical museums. Therefore, we hypothesize that applying these design insights and strategies to physical museums may have a better effect. Overall, we believe that in the near future, as immersive reality technologies become more prevalent in museums, gaze-based audio commentaries will become an important modality, offering significant value for museum experience and learning.

Notes

1. <https://userexperienceawards.com/2017-submissions/aros-art-museum-aros-public>.
2. <https://mw18.mwconf.org/glami/gaze-tracker>.
3. <https://www.mleuven.be/en/eye-tracking-research-m>.
4. <https://zh.wikipedia.org/wiki/%E6%97%A9%E6%98%A5%E5%9C%96>.
5. <https://zh.wikipedia.org/wiki/%E8%B0%BF%E5%B1%B1%E8%A1%8C%E6%97%85%E5%9C%96>.
6. <https://zh.wikipedia.org/wiki/%E5%8C%A1%E5%BB%AC%E5%9C%96>.
7. <https://zh.wikipedia.org/wiki/%E6%BA%AA%E5%B2%B8%E5%9B%BE>.
8. <https://ttsmaker.com/>.
9. <http://www.chinaknowledge.de/Literature/Science/linguangaozhi.html>.
10. https://en.wikipedia.org/wiki/Latin_square.

Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments, and all the participants for actively participating in this experiment.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work is supported by the National Social Science Fund of China [19ZDA046] and the National Natural Science Foundation of China [52205290].

ORCID

Xin Ge  <http://orcid.org/0000-0002-9003-5121>

Xiaoteng Tang  <http://orcid.org/0000-0003-1859-8488>

References

Ahmann, J. S., & Glock, M. D. (1981). *Evaluating student progress: Principles of tests and measurements*. Allyn and Bacon.

- Al-Thani, L. K., & Liginlal, D. (2018). A study of natural interactions with digital Heritage Artifacts. 2018 3rd Digital Heritage International Congress (DigitalHERITAGE) Held Jointly with 2018 24th International Conference on Virtual Systems & Multimedia (VSMM 2018), 1–4. <https://doi.org/10.1109/DigitalHeritage.2018.8810048>
- Ardissono, L., Kuflik, T., & Petrelli, D. (2012). Personalization in cultural heritage: The road travelled and the one ahead. *User Modeling and User-Adapted Interaction*, 22(1-2), 73–99. <https://doi.org/10.1007/s11257-011-9104-x>
- Baradaran Rahimi, F., Boyd, J. E., Eiserman, J. R., Levy, R. M., & Kim, B. (2022). Museum beyond physical walls: An exploration of virtual reality-enhanced experience in an exhibition-like space. *Virtual Reality*, 26(4), 1471–1488. <https://doi.org/10.1007/s10055-022-00643-5>
- Bauer-Krösbacher, C. (2013). Mobile interpretation at cultural attractions: Insights into users and non-users of audio-guides. *Cultural Tourism*, 64–73. <https://doi.org/10.1079/9781845939236.0064>
- Black, G. (2012). *The engaging museum: Developing museums for visitor involvement*. Routledge. <https://doi.org/10.4324/9780203559277>
- Blignaut, P. (2009). Fixation identification: The optimum threshold for a dispersion algorithm. *Attention, Perception & Psychophysics*, 71(4), 881–895. <https://doi.org/10.3758/APP.71.4.881>
- Capece, S., Chivàran, C., Giugliano, G., Laudante, E., Nappi, M. L., & Buono, M. (2024). Advanced systems and technologies for the enhancement of user experience in cultural spaces: An overview. *Heritage Science*, 12(1), 71. <https://doi.org/10.1186/s40494-024-01186-5>
- Chen, C.-C., Kang, X., Li, X.-Z., & Kang, J. (2024). Design and evaluation for improving lantern culture learning experience with augmented reality. *International Journal of Human-Computer Interaction*, 40(6), 1465–1478. <https://doi.org/10.1080/10447318.2023.2193513>
- Chen, X., Ge, X., Wu, Q., & Wang, X. (2024). Actively viewing the audio-visual media: An eye-controlled application for experience and learning in traditional Chinese Painting Exhibitions. *International Journal of Human-Computer Interaction*, 1–29. <https://doi.org/10.1080/10447318.2024.2371691>
- Cheng, M. (2024). From Ink to Pixels: A study on the fusion of traditional Chinese landscape painting and digital media art. *Proceedings of EVA London 2024*. <https://doi.org/10.14236/ewic/EVA2024.19>
- Chiu, T.-W., & Camilli, G. (2013). Comment on 3PL IRT adjustment for guessing. *Applied Psychological Measurement*, 37(1), 76–86. <https://doi.org/10.1177/0146621612459369>
- Chun, M. M., & Turk-Browne, N. B. (2007). Interactions between attention and memory. *Current Opinion in Neurobiology*, 17(2), 177–184. <https://doi.org/10.1016/j.conb.2007.03.005>
- Clay, V., König, P., & König, S. U. (2019). Eye tracking in virtual reality. *Journal of Eye Movement Research*, 12(1), 1. <https://doi.org/10.16910/jemr.12.1.3>
- Clemenson, G. D., Wang, L., Mao, Z., Stark, S. M., & Stark, C. E. L. (2020). Exploring the spatial relationships between real and virtual experiences: What transfers and what doesn't. *Frontiers in Virtual Reality*, 1. <https://doi.org/10.3389/frvir.2020.572122>
- Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research*. SAGE Publications.
- Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches*. SAGE Publications.
- Cristina, S., & Camilleri, K. P. (2018). Unobtrusive and pervasive video-based eye-gaze tracking. *Image and Vision Computing*, 74, 21–40. <https://doi.org/10.1016/j.imavis.2018.04.002>
- Dondi, P., & Porta, M. (2023). Gaze-based human-Computer interaction for museums and exhibitions: Technologies, applications and future perspectives. *Electronics*, 12(14), 3064. <https://doi.org/10.3390/electronics12143064>
- Dondi, P., Porta, M., Donvito, A., & Volpe, G. (2022). A gaze-based interactive system to explore artwork imagery. *Journal on Multimodal User Interfaces*, 16(1), 55–67. <https://doi.org/10.1007/s12193-021-00373-z>
- Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of Educational measurement*. Prentice Hall.
- Falk, J. H. (2009). *Identity and the museum visitor experience*. Routledge. <https://doi.org/10.4324/9781315427058>

- Falk, J. H., & Dierking, L. D. (1997). School field trips: Assessing their long-term impact. *Curator: The Museum Journal*, 40(3), 211–218. <https://doi.org/10.1111/j.2151-6952.1997.tb01304.x>
- Falk, J. H., & Dierking, L. D. (2000). *Learning from museums: Visitor experiences and the making of meaning*. Altamira Press.
- Fan, H., & Poole, M. S. (2006). What is personalization? Perspectives on the design and implementation of personalization in information systems. *Journal of Organizational Computing and Electronic Commerce*, 16(3), 179–202. <https://doi.org/10.1080/10919392.2006.9681199>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>
- Fraenkel, J. R., & Wallen, N. E. (1990). *How to design and evaluate research in education*. Order Department, McGraw Hill Publishing Co.
- Gano, S., & Kinzler, R. (2011). Bringing the museum into the classroom. *Science*, 331(6020), 1028–1029. <https://doi.org/10.1126/science.1197076>
- George, C., Demmler, M., & Hussmann, H. (2018). *Intelligent interruptions for IVR: Investigating the interplay between presence, workload and attention* [Paper presentation]. Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, 1–6. <https://doi.org/10.1145/3170427.3188686>
- Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(9), 904–908. <https://doi.org/10.1177/154193120605000909>
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Advances in psychology*. (Vol. 52, pp. 139–183). [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- Hejtmanek, L., Starrett, M., Ferrer, E., & Ekstrom, A. D. (2020). How much of what we learn in virtual reality transfers to real-world navigation? *Multisensory Research*, 33(4-5), 479–503. <https://doi.org/10.1163/22134808-20201445>
- Helmert, J. R., Pannasch, S., & Velichkovsky, B. M. (2008). Influences of dwell time and cursor control on the performance in gaze driven typing. *Journal of Eye Movement Research*, 2(4), 3. <https://doi.org/10.16910/jemr.2.4.3>
- Hou, W., & Chen, X. (2021). Comparison of eye-based and controller-based selection in virtual reality. *International Journal of Human-Computer Interaction*, 37(5), 484–495. <https://doi.org/10.1080/10447318.2020.1826190>
- Hürst, W., de Coninck, F., & Tan, X. J. (2016). Complementing artworks to create immersive VR museum experiences. *Proceedings of the 13th International Conference on Advances in Computer Entertainment Technology*, 1–6. <https://doi.org/10.1145/3001773.3001806>
- Hutchinson, R., & Eardley, A. F. (2021). Inclusive museum audio guides: ‘Guided looking’ through audio description enhances memorability of artworks for sighted audiences. *Museum Management and Curatorship*, 36(4), 427–446. <https://doi.org/10.1080/09647775.2021.1891563>
- Hutchinson, R., & Eardley, A. F. (2024). ‘I felt I was right there with them’: The impact of sound-enriched audio description on experiencing and remembering artworks, for blind and sighted museum audiences. *Museum Management and Curatorship*, 39(6), 733–750. <https://doi.org/10.1080/09647775.2023.2188482>
- Innocente, C., Ulrich, L., Moos, S., & Vezzetti, E. (2023). A framework study on the use of immersive XR technologies in the cultural heritage domain. *Journal of Cultural Heritage*, 62, 268–283. <https://doi.org/10.1016/j.culher.2023.06.001>
- Ivankova, N. V., Creswell, J. W., & Stick, S. L. (2006). Using mixed-methods sequential explanatory design: From theory to practice. *Field Methods*, 18(1), 3–20. <https://doi.org/10.1177/1525822X05282260>
- Jacob, R. J. K. (1990). What you look at is what you get: Eye movement-based interaction techniques. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 11–18. <https://doi.org/10.1145/97243.97246>
- Jacob, R. J. K. (1991). The use of eye movements in human-computer interaction techniques: What you look at is what you get. *ACM Transactions on Information Systems*, 9(2), 152–169. <https://doi.org/10.1145/123078.128728>
- Jacob, R. J. K. (1993). Eye movement-based human-computer interaction techniques: Toward non-command interfaces. *Advances in Human-Computer Interaction*, 4, 151–190.
- Jarrier, E., & Bourgeon-Renault, D. (2012). Impact of mediation devices on the museum visit experience and on visitors’ behavioural intentions. *International Journal of Arts Management*, 15(1), 18–29.
- Jelavić, Ž., Brezinščak, R., & Škarić, M. (Eds.) (2012). *Old questions, new answers: Quality criteria for museum education*. ICOM.
- Jiménez-Hurtado, C., & Soler Gallego, S. (2015). Museum accessibility through translation: A corpus study of pictorial audio description. In *Audiovisual translation: Taking stock* (pp. 279–298).
- Jin, S., Fan, M., & Kadir, A. (2022). Immersive Spring Morning in the Han Palac e: Learning traditional chinese art via virtual reality and multi-touch Tabletop. *International Journal of Human-Computer Interaction*, 38(3), 213–226. <https://doi.org/10.1080/10447318.2021.1930389>
- Kaghat, F. Z., Azough, A., Fakhour, M., & Meknassi, M. (2020). A new audio augmented reality interaction and adaptation model for museum visits. *Computers & Electrical Engineering*, 84, 106606. <https://doi.org/10.1016/j.compeleceng.2020.106606>
- Kaghat, F. Z., & Cubaud, P. (2010). Fluid interaction in audio-guided museum visit: Authoring tool and visitor device. *Proceedings of the 11th International Conference on Virtual Reality, Archaeology and Cultural Heritage*, 163–170.
- Kang, M. J., Hsu, M., Krajbich, I. M., Loewenstein, G. F., McClure, S. M., Wang, J. T., & Camerer, C. F. (2008). The wick in the candle of learning: Epistemic curiosity activates reward circuitry and enhances memory. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.1308286>
- Kennedy, A. A. U., Thacker, I., Nye, B. D., Sinatra, G. M., Swartout, W., & Lindsey, E. (2021). Promoting interest, positive emotions, and knowledge using augmented reality in a museum setting. *International Journal of Science Education, Part B*, 11(3), 242–258. <https://doi.org/10.1080/21548455.2021.1946619>
- Kim, R. S., Seitz, A. R., & Shams, L. (2008). Benefits of stimulus congruency for multisensory facilitation of visual learning. *PLoS One*, 3(1), e1532. <https://doi.org/10.1371/journal.pone.0001532>
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Krishna, G. (2015). *The best interface is no interface: The simple path to brilliant technology*. Pearson Education.
- Krukar, J., & Dalton, R. C. (2020). How the visitors’ cognitive engagement is driven (but Not Dictated) by the visibility and co-visibility of art exhibits. *Frontiers in Psychology*, 11, 350. <https://doi.org/10.3389/fpsyg.2020.00350>
- Kumar, C., Menges, R., & Staab, S. (2016). Eye-controlled interfaces for multimedia interaction. *IEEE Multimedia*, 23(4), 6–13. <https://doi.org/10.1109/MMUL.2016.52>
- Kuo, Y.-T., Garcia Bravo, E., Whittinghill, D. M., & Kuo, Y.-C. (2024). Walking into a modern painting: The impacts of using virtual reality on student learning performance and experiences in art appreciation. *International Journal of Human-Computer Interaction*, 40(23), 8180–8201. <https://doi.org/10.1080/10447318.2023.2278929>
- Kwok, T. C. K., Kiefer, P., Schinazi, V. R., Adams, B., & Raubal, M. (2019). Gaze-guided narratives: Adapting audio guide content to gaze in virtual and real environments. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. <https://doi.org/10.1145/3290605.3300721>
- Lee, S. J. (2017). A review of audio guides in the era of smart tourism. *Information Systems Frontiers*, 19(4), 705–715. <https://doi.org/10.1007/s10796-016-9666-6>
- Llanes-Jurado, J., Marín-Morales, J., Guixeres, J., & Alcañiz, M. (2020). Development and calibration of an eye-tracking fixation identification algorithm for immersive virtual reality. *Sensors*, 20(17), 4956. <https://doi.org/10.3390/s20174956>
- Machidon, O. M., Duguleana, M., & Carrozzino, M. (2018). Virtual humans in cultural heritage ICT applications: A review. *Journal of*

- Cultural Heritage*, 33, 249–260. <https://doi.org/10.1016/j.culher.2018.01.007>
- Magee, J., Felzer, T., & MacKenzie, I. S. (2015). Camera Mouse + ClickerAID: Dwell vs. single-muscle click actuation in mouse-replacement interfaces. In M. Antona & C. Stephanidis (Eds.), *Universal access in human-computer interaction. Access to today's technologies*. (pp. 74–84). Springer International Publishing. https://doi.org/10.1007/978-3-319-20678-3_8
- Majaranta, P., & Bulling, A. (2014). Eye tracking and eye-based human–computer interaction. In S. H. Fairclough & K. Gilleade (Eds.), *Advances in physiological computing* (pp. 39–65). Springer London. https://doi.org/10.1007/978-1-4471-6392-3_3
- Mayer, R. E. (2005). *The Cambridge handbook of multimedia learning*. Cambridge University Press.
- Mokatren, M., & Kuflik, T. (2016). Exploring the potential contribution of mobile eye-tracking technology in enhancing the museum visit experience. *CEUR Workshop Proceedings*, 1621, 23–31.
- Mokatren, M., Kuflik, T., & Shimshoni, I. (2018). Exploring the potential of a mobile eye tracker as an intuitive indoor pointing device: A case study in cultural heritage. *Future Generation Computer Systems*, 81, 528–541. <https://doi.org/10.1016/j.future.2017.07.007>
- Neale, H., & Nichols, S. (2001). Theme-based content analysis: A flexible method for virtual environment evaluation. *International Journal of Human-Computer Studies*, 55(2), 167–189. <https://doi.org/10.1006/ijhc.2001.0475>
- Niu, Y., Zuo, H., Yang, X., Xue, C., Peng, N., Zhou, L., Zhou, X., & Jin, T. (2021). Improving accuracy of gaze-control tools: Design recommendations for optimum position, sizes, and spacing of interactive objects. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 31(3), 249–269. <https://doi.org/10.1002/hfm.20884>
- Okolo, C. M., Englert, C. S., Bouck, E. C., Heutsche, A., & Wang, H. (2011). The virtual history museum: Learning U.S. history in diverse eighth grade classrooms. *Remedial and Special Education*, 32(5), 417–428. <https://doi.org/10.1177/0741932510362241>
- Oppermann, R., Specht, M., & Jaceniak, I. (1999). Hippie: A nomadic information system. *Handheld and Ubiquitous Computing*, 330–333. https://doi.org/10.1007/3-540-48157-5_37
- Othman, M. K., Petrie, H., & Power, C. (2011). Engaging visitors in museums with technology: Scales for the measurement of visitor and multimedia guide experience. In P. Campos, N. Graham, J. Jorge, N. Nunes, P. Palanque, & M. Winckler (Eds.), *Human-computer interaction – INTERACT 2011* (Vol. 6949, pp. 92–99). Springer. https://doi.org/10.1007/978-3-642-23768-3_8
- Paivio, A., & Csapo, K. (1969). Concrete image and verbal memory codes. *Journal of Experimental Psychology*, 80(2, Pt.1), 279–285. <https://doi.org/10.1037/h0027273>
- Paivio, A., & Csapo, K. (1973). Picture superiority in free recall: Imagery or dual coding? *Cognitive Psychology*, 5(2), 176–206. [https://doi.org/10.1016/0010-0285\(73\)90032-7](https://doi.org/10.1016/0010-0285(73)90032-7)
- Pallud, J., & Monod, E. (2010). User experience of museum technologies: The phenomenological scales. *European Journal of Information Systems*, 19(5), 562–580. <https://doi.org/10.1057/ejis.2010.37>
- Parry, R. (Ed.). (2013). *Museums in a Digital Age*. Routledge. <https://doi.org/10.4324/9780203716083>
- Paulin Hansen, J., W., Andersen, A., & Roed, P. (1995). Eye-gaze control of multimedia systems. In Y. Anzai, K. Ogawa, & H. Mori (Eds.), *Advances in human factors/ergonomics* (Vol. 20, pp. 37–42). Elsevier. [https://doi.org/10.1016/S0921-2647\(06\)80008-0](https://doi.org/10.1016/S0921-2647(06)80008-0)
- Piening, R., Pfeuffer, K., Esteves, A., Mittermeier, T., Prange, S., Schröder, P., & Alt, F. (2021). Looking for Info: Evaluation of Gaze based information retrieval in augmented reality. In C. Ardito, R. Lanzilotti, A. Malizia, H. Petrie, A. Piccinno, G. Desolda, & K. Inkpen (Eds.), *Human-computer interaction – INTERACT 2021* (Vol. 12932, pp. 544–565). Springer International Publishing. https://doi.org/10.1007/978-3-030-85623-6_32
- Plopski, A., Hirzle, T., Norouzi, N., Qian, L., Bruder, G., & Langlotz, T. (2022). The eye in extended reality: A survey on gaze interaction and eye tracking in head-worn extended reality. *ACM Computing Surveys*, 55(3), 53:1–53:39. <https://doi.org/10.1145/3491207>
- Pujol-Tost, L. (2011). Integrating ICT in exhibitions. *Museum Management and Curatorship*, 26(1), 63–79. <https://doi.org/10.1080/09647775.2011.540127>
- Quian Quiroga, R., & Pedreira, C. (2011). How do we see art: An eye-tracker study. *Frontiers in Human Neuroscience*, 5, 98. <https://doi.org/10.3389/fnhum.2011.00098>
- Raptis, G. E., Kavvetos, G., & Katsini, C. (2021). MuMIA: Multimodal interactions to better understand art contexts. *Applied Sciences*, 11(6), 2695. <https://doi.org/10.3390/app11062695>
- Rashed, M. G., Suzuki, R., Lam, A., Kobayashi, Y., & Kuno, Y. (2015). A vision based guide robot system: Initiating proactive social human robot interaction in museum scenarios. *2015 International Conference on Computer and Information Engineering (ICCIE)*, 5–8. <https://doi.org/10.1109/CCIE.2015.7399316>
- Reinhardt, D., Haesler, S., Hurtienne, J., & Wienrich, C. (2019). *Entropy of controller movements reflects mental workload in virtual reality* [Paper presentation]. 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), 802–808. <https://doi.org/10.1109/VR.2019.8797977>
- Rey, F. B., & Casado-Neira, D. (2013). Participation and technology: perception and public expectations about the use of ICTs in Museums. *Procedia Technology*, 9, 697–704. <https://doi.org/10.1016/j.protcy.2013.12.077>
- Rich, J. (2016). Sound, mobility and landscapes of exhibition: Radio-guided tours at the Science Museum, London, 1960–1964. *Journal of Historical Geography*, 52, 61–73. <https://doi.org/10.1016/j.jhg.2016.02.010>
- Ruhland, K., Peters, C. E., Andrist, S., Badler, J. B., Badler, N. I., Gleicher, M., Mutlu, B., & McDonnell, R. (2015). A review of eye gaze in virtual agents, social robotics and HCI: Behaviour generation, user interaction and perception. *Computer Graphics Forum*, 34(6), 299–326. <https://doi.org/10.1111/cgf.12603>
- Ryan, R. M., & Deci, E. L. (2000a). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25(1), 54–67. <https://doi.org/10.1006/ceps.1999.1020>
- Ryan, R. M., & Deci, E. L. (2000b). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *The American Psychologist*, 55(1), 68–78. <https://doi.org/10.1037/0003-066X.55.1.68>
- Saito, M. (2023). Factors in the presentation method of museum audio guides affecting human appreciation behavior. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 45(45).
- Saito, M., Okudo, T., Yamada, M., & Yamada, S. (2023). Identifying visitor's paintings appreciation for AI audio guide in museums. *Proceedings of the 15th International Conference on Agents and Artificial Intelligence*, 55–64. <https://doi.org/10.5220/0011621500003393>
- Salmouka, F., & Gazi, A. (2021). Mapping sonic practices in museum exhibitions – An overview. In M. Shehade & T. Stylianou-Lambert (Eds.), *Emerging technologies and the digital transformation of museums and heritage sites* (pp. 61–75). Springer International Publishing. https://doi.org/10.1007/978-3-030-83647-4_5
- Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. *Proceedings of the Symposium on Eye Tracking Research & Applications – ETRA '00*, 71–78. <https://doi.org/10.1145/355017.355028>
- Sederberg, K. (2013). Bringing the museum into the classroom, and the class into the museum: An approach for content-based instruction. *Die Unterrichtspraxis/Teaching German*, 46(2), 251–262. <https://doi.org/10.1111/tger.10144>
- Seidenari, L., Baecchi, C., Uricchio, T., Ferracani, A., Bertini, M., & Del Bimbo, A. (2017). Deep artwork detection and retrieval for automatic context-aware audio guides. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 13(3s), 1–21. <https://doi.org/10.1145/3092832>
- Serota, N. (1997). *Experience or interpretation: The dilemma of museums of modern art*. Thames and Hudson.
- Shams, L., & Seitz, A. R. (2008). Benefits of multisensory learning. *Trends in Cognitive Sciences*, 12(11), 411–417. <https://doi.org/10.1016/j.tics.2008.07.006>

- Shic, F., Scassellati, B., & Chawarska, K. (2008). The incomplete fixation measure. *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications*, 111–114. <https://doi.org/10.1145/1344471.1344500>
- Steffen, S., Menges, R., Kumar, C., Wechselberger, U., Schaefer, C., & Walber, T. (2017). Schau genau! A Gaze-controlled 3D game for entertainment and education. *Zenodo*. <https://doi.org/10.5281/zenodo.1293424>
- Sugiura, N., Ogura, R., Matsuda, Y., Komuro, T., & Ogawa, K. (2022). Users' content memorization in multi-user interactive public displays. *International Journal of Human-Computer Interaction*, 38(5), 447–455. <https://doi.org/10.1080/10447318.2021.1948686>
- Sun, J. C.-Y., & Yu, S.-J. (2019). Personalized wearable guides or audio guides: An evaluation of personalized museum guides for improving learning achievement and cognitive load. *International Journal of Human-Computer Interaction*, 35(4-5), 404–414. <https://doi.org/10.1080/10447318.2018.1543078>
- Sylaïou, S., & Papaïoannou, G. (2019). ICT in the promotion of arts and cultural heritage education in museums. In A. Kavoura, E. Kefallonitis, & A. Giovanis (Eds.), *Strategic innovative marketing and tourism* (pp. 363–370). Springer International Publishing. https://doi.org/10.1007/978-3-030-12453-3_41
- Tabanfar, R., Chan, H. H. L., Lin, V., Le, T., & Irish, J. C. (2018). Development and face validation of a Virtual Reality Epley Maneuver System (VREMS) for home Epley treatment of benign paroxysmal positional vertigo: A randomized, controlled trial. *American Journal of Otolaryngology*, 39(2), 184–191. <https://doi.org/10.1016/j.amjoto.2017.11.006>
- Thompson, V. A., & Paivio, A. (1994). Memory for pictures and sounds: Independence of auditory and visual codes. *Canadian Journal of Experimental Psychology = Revue Canadienne de Psychologie Experimentale*, 48(3), 380–398. <https://doi.org/10.1037/1196-1961.48.3.380>
- Toyama, T., Kieninger, T., Shafait, F., & Dengel, A. (2011). Museum Guide 2.0 – An Eye-Tracking based Personal Assistant for Museums and Exhibits. In *Proceedings of the International Conference "Re-Thinking Technology in Museums 2011: Emerging Experiences"*, 1–10.
- Vallez, N., Krauss, S., Espinosa-Aranda, J. L., Pagani, A., Seirafi, K., & Deniz, O. (2020). Automatic Museum Audio Guide. *Sensors*, 20(3), 779. <https://doi.org/10.3390/s20030779>
- Velentza, A.-M., Heinke, D., & Wyatt, J. (2020). Museum robot guides or conventional audio guides? An experimental study. *Advanced Robotics*, 34(24), 1571–1580. <https://doi.org/10.1080/01691864.2020.1854113>
- Villani, D., Morganti, F., Cipresso, P., Ruggi, S., Riva, G., & Gilli, G. (2015). Visual exploration patterns of human figures in action: An eye tracker study with art paintings. *Frontiers in Psychology*, 6, 1636. <https://doi.org/10.3389/fpsyg.2015.01636>
- Waern, A., Løvlie, A., Bacon, K., Mathias, N., Eklund, L., Rajkowska, P., Spence, J., Ryding, K., Mortensen, C. H., Olesen, A. R., Malde, S., Darzentas, D., Bodiaj, E., Tennent, P., Martindale, S., Cameron, H., Spors, V., Benford, S., & Back, J. (2022). *Hybrid Museum Experiences: Theory and Design*. Amsterdam University Press. <https://doi.org/10.5117/9789463726443>
- Widdel, H. (1984). Operational problems in analysing eye movements. In *Advances in psychology*. (Vol. 22, pp. 21–29). North-Holland. [https://doi.org/10.1016/S0166-4115\(08\)61814-2](https://doi.org/10.1016/S0166-4115(08)61814-2)
- Winkel, G. H. (1985). Ecological validity issues in field research settings. In *Advances in environmental psychology* (Vol. 5, pp.1–41). Routledge.
- Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., & Wesslén, A. (2012). *Experimentation in software engineering*. Springer Science & Business Media.
- Wollheim, R. (2023). *Painting as an art*. Princeton University Press.
- Wooding, D. S., Mugglestone, M. D., Purdy, K. J., & Gale, A. G. (2002). Eye movements of large populations: I. Implementation and performance of an autonomous public eye tracker. *Behavior Research Methods, Instruments, & Computers*, 34(4), 509–517. <https://doi.org/10.3758/BF03195480>
- Yang, J., & Chan, C. (2019). *Audio-Augmented Museum Experiences with Gaze Tracking* [Paper presentation]. <https://doi.org/10.1145/3365610.3368415>
- Yi, J. H., & Kim, H. S. (2021). User experience research, experience design, and evaluation methods for museum mixed reality experience. *Journal on Computing and Cultural Heritage*, 14(4), 1–28. <https://doi.org/10.1145/3462645>
- Yi, T., Chang, M., Hong, S., & Lee, J.-H. (2021). Use of eye-tracking in artworks to understand information needs of visitors. *International Journal of Human-Computer Interaction*, 37(3), 220–233. <https://doi.org/10.1080/10447318.2020.1818457>
- Yılmaz, K. T., Meral, E., & Başcı Namlı, Z. (2024). A systematic review of the pedagogical roles of technology in ICT-assisted museum learning studies. *Education and Information Technologies*, 29(8), 10069–10103. <https://doi.org/10.1007/s10639-023-12208-3>
- Yuan, C., & Yun, Z. (2016). *Tunable, a VR reconstruction of "Listening to a Guqin" from emperor Zhao Ji* [Paper presentation]. SIGGRAPH ASIA 2016 VR Showcase, 1–2. <https://doi.org/10.1145/2996376.2996379>
- Zhang, X. (2013). The I Don't know option in the vocabulary size test. *TESOL Quarterly*, 47(4), 790–811. <https://doi.org/10.1002/tesq.98>
- Zhou, Y., Chen, J., & Wang, M. (2022). A meta-analytic review on incorporating virtual and augmented reality in museum learning. *Educational Research Review*, 36, 100454. <https://doi.org/10.1016/j.edurev.2022.100454>
- Zimmermann, A., & Lorenz, A. (2008). LISTEN: A user-adaptive audio augmented museum guide. *User Modeling and User-Adapted Interaction*, 18(5), 389–416. <https://doi.org/10.1007/s11257-008-9049-x>

About the authors

Xin Ge is a Phd candidate in design science at Zhejiang University. His research interests are in the design of immersive experience, human-computer interaction within traditional museums, and cultural heritage dissemination (especially traditional Chinese painting).

Xiaojiang Chen is a Researcher of the "Hundred Talents Program" of Zhejiang University. Her research interests are to develop human-computer interaction interfaces, user experience design practices, and related ergonomics experimental cognitive evaluation research.

Xiaoteng Tang is a Phd candidate in design science at Zhejiang University. His research interests include applying cognitive psychology to human-computer interactions and exploring how to use design methods to improve user experience.

Xiaosong Wang is a Professor in the design department at Zhejiang University. His main research interests fall in the broad area of visual experience design, focusing on the use of VR/AR, immersive experience, and other technologies in different scenarios.