Combining Vision language models and gaze tracking in VR art exhibitions

TESSA LIMBEEK, University of Twente, The Netherlands

In recent years, virtual reality and the integration of conversational agents have become more prevalent in art exhibitions, which has led to the exploration of new methods to enhance user experience. Although eye tracking and conversational agents have been utilised separately until now, their combined application remains under-explored. This study investigates how integrating real-time gaze data with a vision language model (VLM) supports the automatic identification of areas of interest (AOIs) and influences user experience in a VR art gallery. To evaluate AOI identification, we first assessed the baseline of the VLM by manually checking the capacity of the VLM with and without contextual information. Afterwards, we introduced the contextual information and compared the answers between the manually defined AOI agent and the gaze-driven agent. To assess the user experience, a user study with 27 participants assessed enjoyment, engagement, personalisation, collaboration, and gaze awareness through a VR visit, a questionnaire and an interview. The results suggest that the VLM, when provided with gaze data on a image and a basic text prompt, can identify AOIs in most cases with a quantitative success rate of 72% of AOIs correctly identified in 53 coordinates. The system can detect AOIs beyond those predefined in the contextual information, generating more focused and relevant responses. Although no statistically significant differences in user experience were observed between gaze-informed and manually guided agents, the findings suggest that a gaze-based approach could support similarly effective user interactions while reducing the manual effort required to define AOIs. This work contributes to the development of adaptive and scalable systems for personalised experiences in VR art environments.

Additional Key Words and Phrases: Virtual Reality, Vision Language Model, VLM, Conversational Agent, Art Exhibition, Cultural Heritage, Eye Gaze Tracking

1 INTRODUCTION

The use of VR has increased in recent years, offering new possibilities for interactive art experiences. Virtual Reality (VR) can be defined as a *"computer-generated digital environment that can be experienced and interacted with as if that environment were real"* [11]. This allows museums to engage visitors in a deeper and meaningful way [12, 25, 28].

Eye gaze tracking, which is a technique to detect and measure eye movements, is used in many domains such as psychology [17, 21] and human-computer interaction (HCI) [5, 6] among others. In the context of VR, gaze tracking can offer valuable insights into user attention and engagement [19]. This study focuses on leveraging gaze tracking in VR to enhance interactive art experiences.

Vision language models (VLM), also referred to as Multimodal language models, are models that can learn simultaneously from images and texts [14]. Combined with eye gaze tracking, they could provide a personalised interaction in a virtual art environment [9].

TScIT 43, July 4, 2025, Enschede, The Netherlands

Although previous studies have highlighted the advantages of eye-tracking in virtual environments [2, 18], its use alongside conversational agents (CA), which serve as a virtual guide that can interact in real-time conversations, has yet to be thoroughly investigated. For instance, Javdani Rikhtehgar et al. [10] investigates the effects of varying levels of gaze awareness on user experience, and concludes that a CA that can tailor its response based on specific areas of interest (AOI) enhances enjoyment. However, these AOIs need to be defined manually, which can come at the cost of precision (e.g. overlapping bounding boxes of AOIs) and requires manual labour. This study seeks to build on that foundation by examining how combining real-time gaze data and images of the paintings in a VLM can support dynamic AOI identification and generate adaptive conversational responses. Specifically, it aims to analyse how a VLM can leverage real-time eye-tracking data to identify a user's areas of interest and how this impacts the user experience, focusing on metrics such as enjoyment, engagement, personalisation, collaboration and gaze-awareness.

This leads to the following research question **(RQ)**:

To what extent does the integration of real-time eye gaze data into a vision language model influence user experience and enhance the identification of users' areas of interest in a virtual art environment?

The **RQ** can be divided into the following sub research questions:

SRQ1: To what extent can a vision language model identify areas of interest in virtual reality paintings using real-time gaze data?

SRQ2: To what extent does the integration of real-time gaze-based AOI detection enhance the system's ability to accurately and informatively identify user-relevant regions?

SRQ3: To what extent does the incorporation of real-time gaze-based AOI detection influence users' experiences in a virtual art environment, particularly with regard to enjoyment, engagement, perceived personalisation, collaboration, and gaze awareness?

2 RELATED WORK

This section reviews the related work relevant to our research. It begins with gaze tracking and its applications, followed by the use of VR in museums. Next, it explores gaze tracking within VR environments and behaviour tracking. The discussion then shifts to multimodal interaction using eye gaze tracking. Finally, the section outlines the identified research gap and presents our contribution.

2.1 Gaze tracking and its applications

Gaze tracking identifies the user's gaze points and the corresponding coordinates. Applications for this technology can be found across multiple domains. These domains include psychology, where it helps

in the analysis of the emotional and cognitive processes [17, 21] and education, where gaze tracking is used to understand learning patterns and improve teaching methods [1, 26]. Other domains include neuroscience to analyse behaviour and study neural responses to visual stimuli [20] and human-computer interaction (HCI) to enhance engagement and design a more personalised experience [5, 6].

2.2 VR in museums

In recent years, the implementation of VR in museums has been researched. One study in particular implemented a virtual museum with an outdoor environment to increase the emotion evoked by cultural heritage [32], while other studies analyse how to create an immersive experience in a virtual art exhibition [15, 30].

2.3 Gaze tracking in VR and behaviour tracking

Since the rise of VR, applications using gaze tracking in virtual reality have become more accessible and frequent. Non-verbal signals, particularly eye contact, are crucial in our everyday interactions as a sign of engagement and as a way to share information [16]. Gaze tracking offers many possibilities in analysing human cognition and behaviour. For instance, Rahman et al. [22] studied students' activities and behaviour patterns to create better teacher-guided VR applications. The study of Mu et al. [19] analyses the correlation between user engagement and levels of interest and found that participants exhibited varying eye gaze patterns, with longer gazes indicating closer inspection and interest in specific artwork details.

2.4 Multimodal interaction with eye gaze tracking

Multimodal interaction facilitates users' engagement with the system through multiple input modes (vision, speech, or touch). It has been demonstrated to enhance user experience, provide flexibility in interaction methods, and accommodate a wide range of user preferences [29]. Eye tracking has been integrated into multimodal applications, notably in cultural heritage systems. For instance, eye tracking has produced museum guides to provide specific information when visitors view specific objects [23, 27]. These studies explore the integration of gaze tracking in multimodal systems but only offer a predefined set of information when requested. Javdani Rikhtehgar et al. [10] explores the integration of a Large Language Model (LLM) with gaze tracking within a VR art environment to enhance user interaction with a more personalised collaboration. Ho et al. [8] explored the integration of a VLM in a virtual art exhibition and found that it creates more diverse and deeper interactions. Moreover, the study of Yan et al. [31] aligns VLM with gaze attention by training an AI model and comparing the results to other models; however, it does not focus on what impact this can have on the user experience.

2.5 Research Gap and Contribution

Prior research by Javdani Rikhtegar et al. [10], which studied the effect of varying levels of eye gaze tracking, demonstrated that conversational agents that can tailor responses based on specific AOIs enhance the user's sense of enjoyment. However, this approach has not yet been extensively explored within the context of VR art environments, as the AOIs need to be defined manually, which is cost-effective. This study aims to replace manually identified AOIs with automatic AOIs to investigate how the integration of real-time gaze data into a VLM impacts the user experience. By doing so, this research seeks to inform the development of more personalised and engaging user experiences in VR art settings.

3 METHODOLOGY

This section outlines the methodology employed in this research, starting with an overview of the system. It then details the procedures used for AOI identification, the design of the user study, and the evaluation measures applied to assess user experience.

3.1 System Overview

Figure 1 illustrates an overview of the current system, the system is extended upon the research of Javdani Rikhtehgar et al. [10]. This system consists of a virtual exhibition developed in Unity[1], where users can explore 3D representations of five artworks and access contextual information through a conversational virtual guide powered by AI. The conversational agent leverages the VLM OpenAI GPT-4o mini[2] and the Rasa[3] framework. The choice for this VLM has been made to remain consistent with previous studies on this topic. To integrate visual data into the conversational agent, we plot the last coordinates on the painting and send the image to the agent.

3.1.1 Integration of real-time gaze data. Participants utilised an HTC VIVE Pro Eye head-mounted display that features built-in eye-tracking capabilities. Gaze data were transmitted in real time on a frame-by-frame basis through the Tobii Pro SDK[4]. The conversational agent functions in two different modes within the same user experience: (1) responsive, where it addresses user inquiries, and (2) proactive, where it starts a conversation after 30 seconds of user inactivity. In both modes, the agent retrieves information from a domain-specific knowledge graph. The proactive prompt is depicted in Figure D.3. To adapt to the context, the agent retrieves the participant's latest gaze coordinates, which are superimposed on the digital visualisation of the artwork (Figure 2). The code snippet below demonstrates how the resulting annotated image and accompanying prompt and possible user input are provided to the VLM. The prompt itself includes relevant metadata about the painting as well as the dialogue history, enabling the model to generate contextually appropriate responses.

```
try:
    messages = [{"role": "system", "content": system_role
        }]
    user_content = []

    if user_text:
        user_content.append({
            "type": "text",
            "text": user_text
        })

    if image_url:
        user_content.append({
            "type": "image_url",
            "image_url": {
```

[1] https://unity.com/
[2] https://openai.com/
[3] https://rasa.com/
[4] https://developer.tobiipro.com/

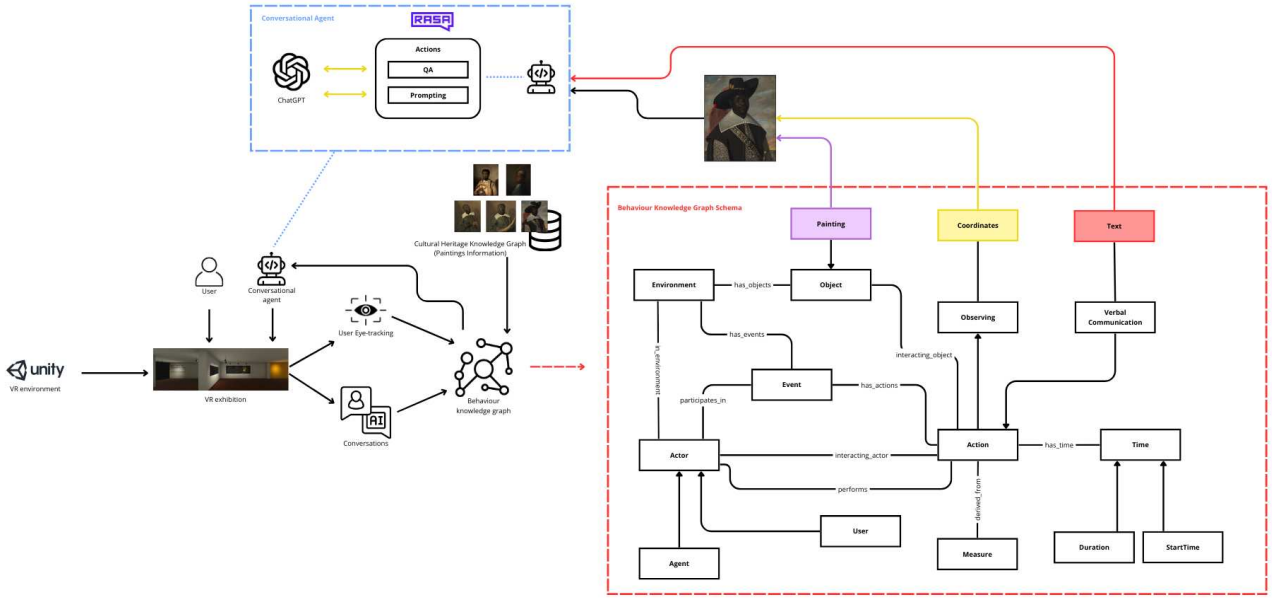


Fig. 1. System Overview and Knowledge Graph Schema

```
                "url": image_url
            }
        })

    # Only add user message if there's content
    if user_content:
        messages.append({
            "role": "user",
            "content": user_content
        })

    completion = client.chat.completions.create(
        model="gpt-4o-mini",
        messages=messages,
    )
```

(a) Example 1

(b) Example 2

Fig. 2. Examples of coordinates plotted onto the paintings

3.2 Methodology for SRQ1: Identification of AOIs

To address this sub-research question, we conducted a two-part comparative evaluation of the agent's effectiveness in identifying user AOIs within VR paintings. In the first part of the evaluation, we first deliberately removed the cultural heritage knowledge graph information about the paintings from the agent's prompt (Figure D.2) to assess how the VLM responds without additional guidance. Approximately ten specific coordinates were selected per painting by the researcher, as presented in Table C.1. The number of coordinates varies between paintings, reflecting differences in the number of identifiable elements across the artworks. The agent's responses to these coordinates were manually checked and categorised into three groups: (1) correctly identified object, (2) incorrectly identified object, and (3) vague or partially correct identification. This allowed us to assess the VLM's baseline interpretive capabilities in the absence of background knowledge. Subsequently, to access the effect of the knowledge graph, the same coordinates from The African King Caspar were re-evaluated using a prompt that included the cultural heritage knowledge graph (Figure D.3). The agent's responses were again categorised using the same criteria, allowing for a direct comparison to assess the impact of contextual information on AOI identification. An overview of these coordinates is presented in Table C.2.

3.3 Methodology for SRQ2: Effect of gaze-data on the identification of AOIs

To address this sub-research question, we evaluated the impact of gaze tracking on AOI identification by comparing two agent configurations: one using manually defined AOIs and the other using automatically inferred AOIs based on real-time gaze data. In both cases, historical and contextual information was integrated into the prompts (Figure D.1 for the manual-AOI agent and Figure D.3 for the gaze-driven agent). As shown in Table C.3, specific coordinates were again selected by the researcher across the various paintings, and corresponding nodes were created in the knowledge graph to represent these AOIs. The agent's performance in each condition was assessed by initiating new conversations using the predefined coordinates, with no dialogue history to avoid bias. The outputs from both conditions for each coordinate were then compared on the accuracy of AOI identification and the depth of contextual information provided in the responses.

3.4 Methodology for SRQ3: Experiment setup

To answer this sub-research question, a user study has been performed. The target group for the participants are university students in the Netherlands aged 17 to 30. The user study has been reviewed by the Ethics Committee of the University of Twente. The study consists of 3 phases: the VR visit, a survey and an interview.

3.4.1 VR visit. For this study, a between-subjects experimental design was employed to compare two conversational agents: the second platform developed by Javdani Rikhtehgar et al. [10], hereafter referred to as the Manual-AOI agent, and the system described in this work, referred to as the gaze-driven agent. The participants participated in the designed system using a VLM connected to gaze data. Each participant interacts with the VR exhibition at their own pace, viewing paintings and reading texts. Participants are able to engage with the virtual agent by asking questions, following its prompts, or simply listening to its commentary. The VR setup will be the same as in the research by Javdani Rikhtehgar et al. [10], containing 5 paintings. This decision was made to enable comparison with these results in order to evaluate the user experience. After the VR visit, the participant completes a survey. Subsequently, the participants share their thoughts in a follow-up interview.

3.4.2 Survey. This research used a survey to gather demographic data from participants and assess the effectiveness of the CA. The survey includes questions regarding participants' gender, their museum visiting behaviours, their familiarity with VR technology and virtual museums, as well as their preferred ways of receiving information about the exhibition. Participants were also asked to indicate the agent's effectiveness using a 5-point Likert scale (from Strongly Disagreed to Strongly Agreed).

3.4.3 Interview. Subsequently, a semi-structured interview has been conducted to gain deeper insights into participants' experiences. The interview begins with questions regarding participants' familiarity with and comfort in VR environments and their overall impressions of the VR experience. After that, participants were asked about which sections they remembered, if they acquired any new knowledge, and whether they found the content beneficial. Lastly, we asked if the experience felt personalised and why, and if they felt that the conversational agent reacted to their AOI.

3.4.4 Analysis. We assessed the user experience using five main criteria: enjoyment, engagement, perceived personalisation, perceived collaboration and gaze awareness. Table B.1 outlines the specific questions associated with each metric. In addition, we examine the duration of time participants spent in the environment, and insights from interview responses to develop a well-rounded understanding of their experiences.

Enjoyment captures users' satisfaction and pleasure during the interaction with the virtual agent and their intention to reuse the system in the future. **Engagement** assesses the degree of emotional and cognitive involvement, using questions related to time perception, attentiveness, and interest in the agent's responses. **Perceived personalisation** measures how well the agent adapts to individual user preferences, based on adaptability and recognition of user interests. **Collaboration** captures the sense of working together toward shared goals, including mutual understanding and responsiveness. **Gaze awareness** evaluates whether users feel the agent detects and reacts to their visual attention, thereby enhancing the sense of interactivity and presence.

These five aspects were assessed through Likert-scale questions, with statistical differences between groups tested using t-tests. In addition, the duration of time spent in the VR environment was used as a behavioural indicator of engagement. Interview data further enriches the evaluation by capturing user reflections, preferences, and suggestions, offering qualitative insights

4 RESULTS

This section describes the results. First, we will address the results of SRQ1 which assessed the baseline performance of the VLM, after which we will address the results of SRQ2 where the effect of gaze-data on the identification of AOIs is assesed. Lastly we will cover the results of SRQ3, which evaluates the user experience.

4.1 Results for SRQ1: Identification of AOIs

In this phase, the VLM was prompted with visual input (Figure 2) and a basic prompt with no supplementary background about the artworks. (Figure D.2). Across 53 coordinates, the model correctly identified the object of focus in approximately 72% of the cases, typically visually distinct features such as white collars, buttons, and hair. In 6% of cases, the model produced incorrect identifications, for example, not recognizing the pink bow in Portrait of Dom Miguel de Castro, depicted in Figure 2b, or failing to detect the ring worn by King Caspar, instead referring only to the hand and fingers. The remaining 23% were vague or only partially accurate; for instance, when the coordinate corresponded to the mouth, the agent described the overall facial expression rather than the specific feature:

> *"You're observing a striking portrait that beautifully captures the subject's **expressions** and attire. The intricate details of the jewellery, especially around the neck, reflect richness and cultural significance, while the neutral background enhances the subject's presence ..."*

Similar issues included referencing jewellery generically instead of specifying a necklace. Table C.1 gives a full overview of all the

coordinates tested and their identification. Despite the lack of contextual input, the model indicates a baseline visual comprehension by identifying AOIs beyond those manually defined.

To evaluate the impact of the knowledge graph, we re-tested 14 coordinates from King Caspar's portrait providing the knowledge graph. Both setups yielded similar accuracy, with 57% correct identifications. The agent without the knowledge graph produced incorrect results in 7% of cases, while the agent with the graph erred in 14%. Partial identifications occurred in 35% and 28% of responses, respectively. These findings suggest that, regardless of context, the model can identify AOIs. However, the knowledge graph enhances the depth of the responses and provides knowledge about the object, as is illustrated below, where the gaze is directed at the incense pot of The African King Caspar:

> *"It seems you're focusing on a ..., particularly around the figure's hand and the **pot** they are holding. The portrayal of the hand is quite detailed, showing careful attention to texture and anatomical accuracy, which reflects the **artist's skill** in rendering human forms. **The metallic sheen of the pot** adds contrast and draws the viewer's eye, enhancing the overall **richness** of the composition ...""*
>
> *– Without knowledge graph*
>
> *"You're currently observing a key detail of "The African King Caspar." The figure is holding an **incense pot**, which symbolizes the **gift of frankincense**, representing **Jesus' divinity**. This golden vessel, ornately decorated, emphasizes its importance as one of the three gifts presented to **Jesus by the magi**, and reflects the cultural significance of incense in both religious and social contexts during that time ...""*
>
> *– With knowledge graph*

4.2 Results for SRQ2: Effect of gaze-data on the identification of AOIs

The second phase introduced contextual information into the prompt, including historical and thematic details. Two agent configurations were compared: the first relied on approximately five manually defined AOIs per painting, while the second used real-time gaze tracking to dynamically infer AOIs. Both had access to the same contextual knowledge base. An overview of the findings can be found in Table C.3.

The gaze-driven agent often identified AOIs beyond those described in the prompt, while the manual-AOI agent focused mainly on the predefined areas. For instance, in Head of a Boy in a Turban, the gaze-driven agent correctly identified the golden embellishments on the blue garment, whereas the manual agent provided a description of the garment. A similar case appeared in Portrait of Dom Miguel de Castro (Figure 2b), where the gaze agent described the ornate belt, while the manual agent referenced the garment more generally.

The gaze-driven agent also indicated greater spatial precision. In The African King Caspar, the necklace AOI did not fully cover the object, yet the gaze-driven agent recognized a nearby gaze point as part of it, offering a more complete description. Additionally, it also frequently incorporated interpretive elements, such as emotional expression or social status, when the gaze was directed at the subject's face or eyes; for example:

> *"It looks like you're focusing on the face of King Caspar... **The expression is one of pride and confidence**, which signifies his high status..."*

Both agents performed similarly with prominent features like the turban or Diego Bemba's box. Interestingly, the manual-AOI agent provided richer background descriptions when AOIs were located in those areas, often referencing visual techniques. The gaze-driven agent responded more generally to background-focused gaze, unless it landed close to another object, at which point it often shifted and delivered a more detailed interpretation, reflecting sensitivity to subtle gaze shifts. That said, the gaze-driven agent guided the conversation toward the predefined AOIs, even when the user's gaze was focused elsewhere.

4.3 Results for SRQ3: User experience

This section summarises the findings from the experiment. The results from Group 1 were previously gathered in research by Javdani Rikhtehgar et al. [10], but will be discussed here as well for comparison.

4.3.1 Questionnaire results. This section presents the results gathered from the questionnaire in the user study. The study included 27 participants, with 17 assigned to the manual-AOI condition (Group 1) and 10 to the gaze-driven agent (Group 2). All participants were between the ages of 20 and 30 years old. The sample consisted of 14 males, 11 females, and 2 individuals who chose not to disclose their gender. In terms of familiarity with VR, 1 participant was unfamiliar with it, 8 had heard of it, 12 had experienced it, and 6 were very knowledgeable. Regarding museum attendance, 9 participants visited rarely, 17 participants visited occasionally, and 1 participant visited frequently. Out of the paintings shown in the VR experience, 23 were unfamiliar, 3 were somewhat familiar, and 1 was very familiar. Interest in virtual agents was expressed by 13 participants, while 3 expressed disinterest and 11 were uncertain. For personalised virtual agents, 22 participants showed interest, 2 were not interested, and 3 were uncertain. These results can be visualized in Figure 3 displaying the results in pie charts.

The outcomes of the Likert-scale assessment are presented in Table B.1 and illustrated in Figure 4, which displays box plots that demonstrate the data distribution for each group. To assess the assumption of normality, Shapiro–Wilk tests were conducted for each evaluation measure across the two groups. The corresponding Q-Q plots are presented in the appendix in Figure E.1 and Figure E.2. Results indicate that all variables, except for perceived personalisation in Group 1, did not significantly deviate from normality. In addition, Levene's tests for equality of variances were conducted for each dependent variable, indicating that the assumption of homogeneity of variances was met. Accordingly, independent samples t-tests were conducted to assess between-group differences across all evaluation measures. The analyses indicated no statistically significant differences between experimental conditions on any of the dependent

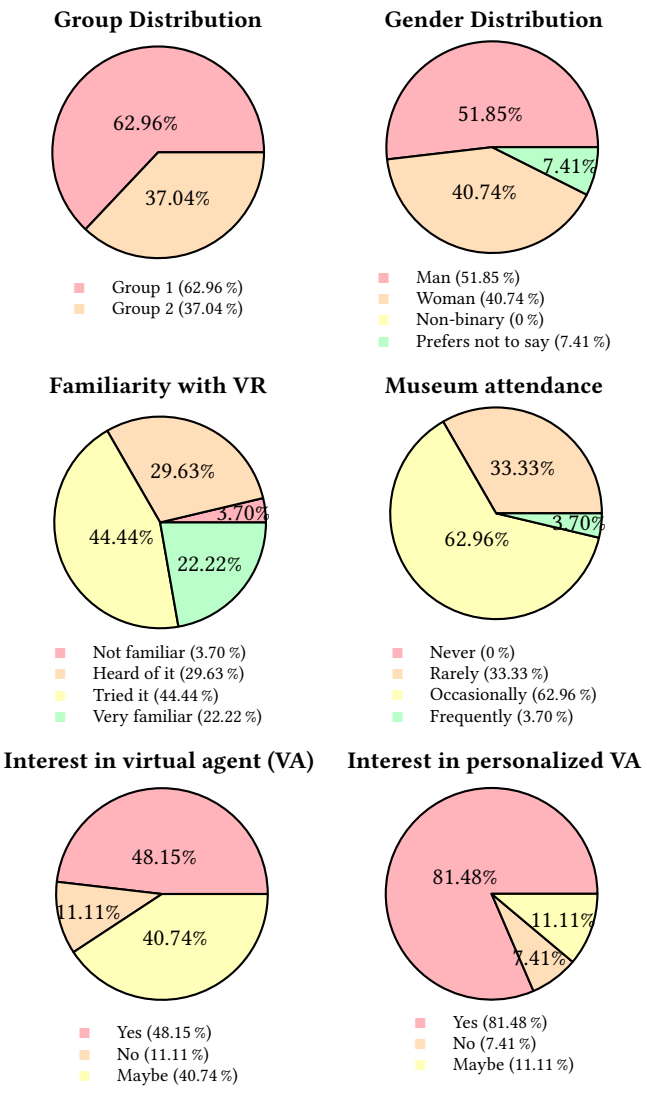


Fig. 3. Pie charts of the Questionnaire results

variables: enjoyment, engagement, gaze awareness, and collaboration (all $p > .05$). For the variable personalisation, a Mann–Whitney U test was used due to a violation of normality assumptions; this analysis also showed no significant difference between conditions. The effect sizes indicated a medium effect for enjoyment (Cohen's $d = 0.693$), a small to medium effect for engagement (Cohen's $d = 0.47$), and small effects for personalisation ($r = 0.13$), collaboration (Cohen's $d = 0.127$), and gaze awareness (Cohen's $d = 0.099$). The average time spent in the VR is 699.19 and 576.7 seconds for Group 1 and Group 2, respectively.

4.3.2 Interview results. The semi-structured interviews conducted with Group 2 revealed several recurring themes related to the user experience in the VR environment. The four primary themes identified were Timing, Personalisation, AOIs and Information, and Speech Input.

Timing. Several participants commented on the timing of interactions within the experience. In some cases, agent responses were perceived as premature, occurring while the user was still observing a painting. One participant noted, *"The response time is of course slow, but that's kind of a given"*. Additionally, overlapping functions specifically, user-initiated questions and automated prompts resulted in extended responses or delays, as one participant notes *"the agent responded by the time you were looking at something else"*. This led to confusion when users had already moved on to a new painting, yet the agent continued discussing the previous one. Some participants also expressed a desire for greater autonomy during the experience, with one remarking that *"I just want to take a look at my own pace"*.

Personalisation. Most participants found the interaction to be personalised and appreciated the ability to ask their own questions. Several users observed that, after several inquiries, the agent appeared to tailor its responses to their interests, for example, focusing on the symbolism of colours or historical context. However, two participants felt the experience lacked sufficient personalisation. This perception stemmed from the agent's tendency to focus on pre-scripted objects, limiting its ability to respond to broader or off-topic queries. One participant expressed disappointment when the agent declined to answer a question outside the museum's content scope. To enhance personalisation, several participants suggested implementing a menu at the beginning of the experience to select preferred topics and delivery modes for information.

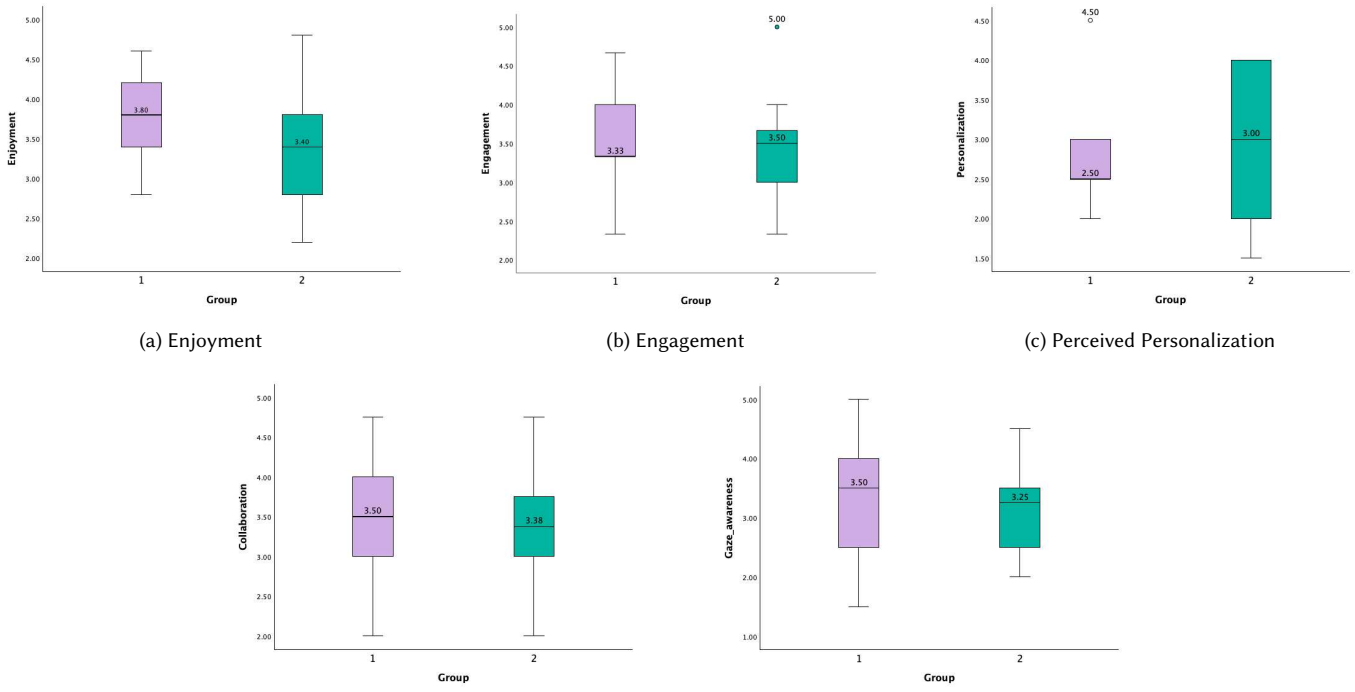
AOIs and Information. Most participants reported that the agent responded in alignment with their interests, and the information provided was generally perceived as useful and engaging. Nonetheless, there were notable exceptions. On some occasions, the agent failed to recognise specific objects in the scene. For instance, in one interaction, a participant inquired about a "pink bow" worn by Dom Miguel de Castro, depicted in Figure 2b, which the agent incorrectly claimed was not present. The agent responded:

> *"The painting you are referencing features Dom Miguel de Castro in an elegantly styled outfit, but there isn't a pink bow depicted in the artwork; instead, his attire includes a cavalier hat, topped with a striking red ostrich feather…"*

This inaccuracy was particularly disappointing for the participant, who expressed an interest in small visual details.

Speech Input. Speech input emerged as a significant usability challenge. Participants frequently encountered difficulties with voice recognition, often needing to repeat their questions. On average, participants reported having to repeat their queries approximately 5.8 times[5] over the entire visit. These issues were attributed to the need to hold the input button for several seconds after finishing a question, as well as to the system's sensitivity to speech pace and phrasing. The system only reliably detected clearly articulated, direct questions and failed to interpret more natural, conversational phrasing where the question emerged mid-sentence.

[5]This value reflects the arithmetic mean calculated from all interactions with the agent during the experience.



(a) Enjoyment (b) Engagement (c) Perceived Personalization

(d) Collaboration (e) Gaze awareness

Fig. 4. Results- Boxplot showing distribution for each group

5 DISCUSSION

5.1 Discussion for SRQ1: Identification of AOIs

The findings from this study indicate that the VLM, both with and without contextual information, shows a capacity for recognizing visually distinctive features in VR paintings. The relatively high accuracy in the first phase suggests the model can produce relevant content from visual input alone. However, vague or incorrect responses reveal the model's limitations, e especially with nuanced or symbolically rich AOIs. In some cases, the mistakes might be due to the small size of the object and it was partly covered by the coordinate marker. While contextual information enhances the informativeness of response, often adding background and interpretive detail, this was tested on only 14 coordinates from a single painting, limiting generalisability. Additionally, a key limitation is the potential bias in coordinate selection, as these were manually chosen by the researcher, potentially favouring certain features or interpretations over others.

5.2 Discussion for SRQ2: Effect of gaze-data on the identification of AOIs

The second part of the study explored the benefits of using real-time gaze data. For this we introduced contextual information as it provided more informative answers. While both agents used the same historical and descriptive context, the gaze-informed agent more often aligned its responses with the user's actual focus. This suggests that gaze tracking can support more adaptive and targeted interactions by guiding the agent toward dynamically relevant content.

However, limitations emerged, the gaze-informed agent sometimes prioritized contextually described AOIs over the actual gaze point. This anchoring bias, introduced by providing more detail on approximately five AOIs per painting, led the agent to steer interactions toward those focal points, even when gaze indicated otherwise, reducing responsiveness and flexibility.

Interestingly, the fact that the manual-AOI agent often offered a more informative response when the gaze landed on the background could suggest a limitation of the VLM in recognizing the background. However, when the gaze was positioned near a foreground object, the gaze-driven agent indicates a nuanced capacity to shift focus and provide detailed interpretations of adjacent visual elements.

These findings suggest that while gaze-informed systems reduce manual effort and increase flexibility, contextual prompts must be carefully designed to avoid overemphasising specific AOIs. Future work could explore prompt strategies that better balance visual and contextual input, such as including the full painting alongside the gaze coordinate to ensure no details are obscured.

5.3 Discussion SQR3: User experience

This section discusses the key findings of the user study, examining how participants experienced the AI-guided VR museum tour.

Timing. One of the central themes was the issue of timing. Several participants expressed a preference for user-initiated prompting, which would allow them greater autonomy and more time to explore

at their own pace. These findings reflect broader concerns in HCI. For example, Schönau [24] highlights that proactive system prompts may lead users to delegate decision-making, thereby diminishing their sense of agency. In this study, participants reported frustration when the agent interrupted them while they were still focused on a painting, indicating a disruption of user autonomy and engagement. Additionally, delayed or excessively long responses from the agent were noted to impair the flow of interaction and reduce immersion. This aligns with the framework of Csikszentmihalyi et al. [4], which notes that poorly timed system behaviours can disrupt the user's sense of presence and engagement.

Personalisation. Personalisation emerged as another significant theme. Most participants appreciated being able to ask their own questions and felt the agent adapted to their interests over time, aligning with findings that personalisation enhances user satisfaction and engagement [3, 7]. However, two participants felt the experience lacked personalisation, noting that the agent often prioritised AOIs with richer predefined information in the knowledge graph. Although the agent could detect more objects visually, the prompt tended to favour AOIs with more contextual data, limiting response diversity and overlooking visually identified elements. This highlights how the prompt generation process may unintentionally constrain content variety, reducing the perceived depth and personalization of the interaction.

AOIs and Information. Information accuracy and the handling of AOIs were also central to the user experience. As noted in the results, the agent occasionally failed to correctly identify visual elements, errors of this nature can undermine user trust [13], particularly for users with a strong interest in visual detail. While many appreciated the agent's attention to specific elements, some preferred broader contextual or historical insights, suggesting a need for adaptive content delivery based on user preferences.

Speech input. Speech input emerged as a further challenge. Participants encountered difficulties with needing to hold down a button for an extended time and articulating their questions slowly. These technical constraints increased cognitive load and reduced the natural interaction. Such interruptions impaired the natural conversational flow and contributed to user frustration. As a result, problems with speech input not only reduced the quality of the interaction but also made the experience feel less immersive. Users felt less in control because they had to adapt their communication style to the system's limitations.

Finally, the results of the Likert-scale questions did not reveal significant differences between the conditions tested. This may suggest that user experience remained relatively consistent across conditions. Notably, this points to the potential of reducing manual labour in design: gaze-based AOI detection can yield experiences comparable to those created manually.

A few limitations of this study should be acknowledged. The relatively small sample sizes (17 and 10 participants) may limit the generalisability of the findings, despite being within the norm for VR research. Additionally, most participants from Group 2 reported that they only occasionally visit museums and were not particularly interested in this type of art exhibition. This may have influenced their level of engagement during the experience and could have affected the depth or variability of their responses.

Future work could explore expanding the knowledge base available to the agent to support a broader range of object-related content. Enhancements to voice recognition accuracy and support for more natural phrasing would likely improve user experience. Providing users with greater control, for example, via a button next to each painting to request information, could support autonomy and allow comparative studies on different interaction modalities. Additionally, future research could investigate how different forms of personalisation, including upfront preference selection, impact engagement and satisfaction.

6 CONCLUSION

This study investigated the effectiveness of a vision language model (VLM) in detecting user areas of interest (AOIs) using real-time gaze data in virtual reality (VR) art environments and examined how this affects the perceived user experience. The research was structured in three parts. The first part focused on evaluating the VLM's ability to identify AOIs based on visual input with and without contextual information. The second part consisted of a comparison of two agent configurations that incorporated contextual information: one guided by manually predefined AOIs and the other by real-time gaze tracking. The third part consisted of a user study designed to assess how the integration of gaze tracking influenced users' experience.

The findings suggest that the VLM can successfully identify AOIs, even without prior information, accurately recognising elements intended by the user in most instances. In the subsequent phase, the agent utilising gaze tracking exhibited an improved capacity to detect user-specific AOIs beyond those manually established, facilitating more precise and responsive interactions. Although some inclination towards predefined content was noted, the gaze-tracking agent provided greater adaptability in matching the user's genuine focus.

There were no statistically significant variations in user experience measurements between the manually guided and gaze-driven scenarios. This comparability may suggest that similar levels of engagement and perceived customisation can be attained without the labour-intensive task of manually defining AOIs. This suggests that real-time gaze tracking may represent a feasible and scalable option for developing adaptive, user-centred interactions in immersive artistic experiences.

Overall, these results suggest the potential of gaze-aware systems to facilitate automatic AOI detection and personalised interaction in VR contexts, thereby minimising the effort required for defining AOIs while still delivering meaningful user experiences.

ACKNOWLEDGMENTS

I would like to thank D. Javdani Rikhtehgar and dr. S. Wang for their guidance and assistance during this research.

REFERENCES

[1] Teresa Busjahn, Carsten Schulte, Bonita Sharif, Simon, Andrew Begel, Michael Hansen, Roman Bednarik, Paul Orlov, Petri Ihantola, Galina Shchekotova, and Maria Antropova. 2014. Eye tracking in computing education. In *Proceedings of the tenth annual conference on International computing education research (ICER*

'14). Association for Computing Machinery, New York, NY, USA, 3–10. https://doi.org/10.1145/2632320.2632344

[2] Lizhou Cao, Huadong Zhang, Chao Peng, and Jeffrey T. Hansberger. 2023. Real-time multimodal interaction in virtual reality - a case study with a large virtual interface. *Multimedia Tools and Applications* 82, 16 (July 2023), 25427–25448. https://doi.org/10.1007/s11042-023-14381-6

[3] Zuen Cen and Yuxin Zhao. 2024. Enhancing User Engagement through Adaptive Interfaces: A Study on Real-time Personalization in Web Applications. *Journal of Economic Theory and Business Management* 1, 6 (Dec. 2024), 1–7. https://doi.org/10.70393/6a6574626d.323332 Number: 6.

[4] Mihaly Csikszentmihalyi, Sami Abuhamdeh, and Jeanne Nakamura. 2014. Flow. In *Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi*, Mihaly Csikszentmihalyi (Ed.). Springer Netherlands, Dordrecht, 227–238. https://doi.org/10.1007/978-94-017-9088-8_15

[5] Dipankar Das, Md. Golam Rashed, Yoshinori Kobayashi, and Yoshinori Kuno. 2015. Supporting Human–Robot Interaction Based on the Level of Visual Focus of Attention. *IEEE Transactions on Human-Machine Systems* 45, 6 (Dec. 2015), 664–675. https://doi.org/10.1109/THMS.2015.2445856

[6] Piercarlo Dondi and Marco Porta. 2023. Gaze-Based Human–Computer Interaction for Museums and Exhibitions: Technologies, Applications and Future Perspectives. *Electronics* 12, 14 (Jan. 2023), 3064. https://doi.org/10.3390/electronics12143064 Number: 14 Publisher: Multidisciplinary Digital Publishing Institute.

[7] Ahmad Heryanto, Yonis Gulzar, and Gene Marck. 2023. A Novel Framework for Enhancing User Experience in Virtual Reality Environments. *International Journal of Computer Engineering in Research Trends* 10, 2 (Feb. 2023), 61–68. https://doi.org/10.22362/ijcert/2023/v10/i02/v10i0203 Number: 2.

[8] Hoang Phuoc Ho, Vani Ramesh, Ivo Zaloudek, Delaram Javdani Rikhtehgar, and Shenghui Wang. 2025. Enhancing Visitor Engagement in Interactive Art Exhibitions with Visual-Enhanced Conversational Agents. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 660–671. https://doi.org/10.1145/3708359.3712145

[9] Delaram Javdani Rikhtehgar, Shenghui Wang, Hester Huitema, Julia Alvares, Stefan Schlobach, Carolien Rieffe, and Dirk Heylen. 2023. Personalizing Cultural Heritage Access in a Virtual Reality Exhibition: A User Study on Viewing Behavior and Content Preferences. In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization (UMAP '23 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 379–387. https://doi.org/10.1145/3563359.3596666

[10] Delaram Javdani Rikhtehgar, Shenghui Wang, Stefan SCHLOBACH, and Dirk Heylen. 2025. Real-Time Gaze Awareness in Conversational Agents: Enhancing Collaboration and Personalization in VR Art Experiences.

[11] Jason Jerald. 2015. *The VR Book: Human-Centered Design for Virtual Reality*. Morgan & Claypool, n.p. https://books.google.nl/books?hl=nl&lr=&id=ZEBiDwAAQBAJ&oi=fnd&pg=PR11&dq=the+VR+book&ots=0Ao8CLudX-&sig=QiKKE-WT9A1u1a_jLp2JlWgcJHc#v=onepage&q=the%20VR%20book&f=false Google-Books-ID: ZEBiDwAAQBAJ.

[12] Timothy Jung, M. Claudia tom Dieck, Hyunae Lee, and Namho Chung. 2016. Effects of Virtual Reality and Augmented Reality on Visitor Experiences in Museum. In *Information and Communication Technologies in Tourism 2016*, Alessandro Inversini and Roland Schegg (Eds.). Springer International Publishing, Cham, 621–635. https://doi.org/10.1007/978-3-319-28231-2_45

[13] Patricia K. Kahr, Gerrit Rooks, Chris Snijders, and Martijn C. Willemsen. 2024. The Trust Recovery Journey. The Effect of Timing of Errors on the Willingness to Follow AI Advice.. In *Proceedings of the 29th International Conference on Intelligent User Interfaces (IUI '24)*. Association for Computing Machinery, New York, NY, USA, 609–622. https://doi.org/10.1145/3640543.3645167

[14] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Leo Tronchon. 2024. Building and better understanding vision-language models: insights and future directions. https://openreview.net/forum?id=iSL0FHZStr

[15] Hyunae Lee, Timothy Hyungsoo Jung, M. Claudia tom Dieck, and Namho Chung. 2020. Experiencing immersive virtual reality in museums. *Information & Management* 57, 5 (July 2020), 103229. https://doi.org/10.1016/j.im.2019.103229

[16] Anatole Lécuyer. 2017. Playing with Senses in VR: Alternate Perceptions Combining Vision and Touch. *IEEE Computer Graphics and Applications* 37, 1 (Jan. 2017), 20–26. https://doi.org/10.1109/MCG.2017.14

[17] Maria Laura Mele and Stefano Federici. 2012. Gaze and eye-tracking solutions for psychological research. *Cognitive Processing* 13, 1 (Aug. 2012), 261–265. https://doi.org/10.1007/s10339-012-0499-z

[18] Jesús Moreno-Arjonilla, Alfonso López-Ruiz, J. Roberto Jiménez-Pérez, José E. Callejas-Aguilera, and Juan M. Jurado. 2024. Eye-tracking on virtual reality: a survey. *Virtual Reality* 28, 1 (Feb. 2024), 38. https://doi.org/10.1007/s10055-023-00903-y

[19] Mu Mu, Murtada Dohan, Alison Goodyear, Gary Hill, Cleyon Johns, and Andreas Mauthe. 2024. User attention and behaviour in virtual reality art encounter. *Multimedia Tools and Applications* 83, 15 (May 2024), 46595–46624. https://doi.org/10.1007/s11042-022-13365-2

[20] Livia Popa, Ovidiu Selejan, Allan Scott, Dafin F. Mureşanu, Maria Balea, and Alexandru Rafila. 2015. Reading beyond the glance: eye tracking in neurosciences. *Neurological Sciences* 36, 5 (May 2015), 683–688. https://doi.org/10.1007/s10072-015-2076-6

[21] Rima-Maria Rahal and Susann Fiedler. 2019. Understanding cognitive and affective mechanisms in social psychology through eye-tracking. *Journal of Experimental Social Psychology* 85 (Nov. 2019), 103842. https://doi.org/10.1016/j.jesp.2019.103842

[22] Yitoshee Rahman, Sarker Monojit Asish, Adil Khokhar, Arun K Kulshreshth, and Christoph W Borst. 2019. Gaze Data Visualizations for Educational VR Applications. In *Symposium on Spatial User Interaction (SUI '19)*. Association for Computing Machinery, New York, NY, USA, 1–2. https://doi.org/10.1145/3357251.3358752

[23] M. Golam Rashed, R. Suzuki, A. Lam, Y. Kobayashi, and Y. Kuno. 2015. A vision based guide robot system: Initiating proactive social human robot interaction in museum scenarios. In *2015 International Conference on Computer and Information Engineering (ICCIE)*. IEEE, Singapore, 5–8. https://doi.org/10.1109/CCIE.2015.7399316

[24] Andreas Schönau. 2023. Agency in augmented reality: exploring the ethics of Facebook's AI-powered predictive recommendation system. *Ai and Ethics* 3, 2 (2023), 407–417. https://doi.org/10.1007/s43681-022-00158-4

[25] Maria Shehade and Theopisti Stylianou-Lambert. 2020. Virtual Reality in Museums: Exploring the Experiences of Museum Professionals. *Applied Sciences* 10, 11 (Jan. 2020), 4031. https://doi.org/10.3390/app10114031 Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

[26] Yuyang Sun, Qingzhong Li, Honggen Zhang, and Jiancheng Zou. 2018. The Application of Eye Tracking in Education. In *Advances in Intelligent Information Hiding and Multimedia Signal Processing*, Jeng-Shyang Pan, Pei-Wei Tsai, Junzo Watada, and Lakhmi C. Jain (Eds.). Springer International Publishing, Cham, 27–33. https://doi.org/10.1007/978-3-319-63859-1_4

[27] Takumi Toyama, Thomas Kieninger, aisal Shafait, and Andreas Dengel. 2011. Museum Guide 2.0 – An Eye-Tracking based Personal Assistant for Museums and Exhibits. In *Proceedings of the International Conference on Re-Thinking Technology in Museums*. University of Limerick, Limerick, Ireland, 1–11. https://www.researchgate.net/profile/Andreas-Dengel/publication/267793771_Museum_Guide_20_-_An_Eye-Tracking_based_Personal_Assistant_for_Museums_and_Exhibits/links/54d0b61e0cf298d65668296f/Museum-Guide-20-An-Eye-Tracking-based-Personal-Assistant-for-Museums-and-Exhibits.pdf

[28] Mariapina Trunfio, Lucia , Maria Della, Campana , Salvatore, , and Adele Magnelli. 2022. Innovating the cultural heritage museum service model through virtual reality and augmented reality: the effects on the overall visitor experience and satisfaction. *Journal of Heritage Tourism* 17, 1 (Jan. 2022), 1–19. https://doi.org/10.1080/1743873X.2020.1850742 Publisher: Routledge _eprint: https://doi.org/10.1080/1743873X.2020.1850742.

[29] Matthew Turk. 2014. Multimodal interaction: A review. *Pattern Recognition Letters* 36 (Jan. 2014), 189–195. https://doi.org/10.1016/j.patrec.2013.07.003

[30] Rafal Wojciechowski, Krzysztof Walczak, Martin White, and Wojciech Cellary. 2004. Building Virtual and Augmented Reality museum exhibitions. In *Proceedings of the ninth international conference on 3D Web technology (Web3D '04)*. Association for Computing Machinery, New York, NY, USA, 135–144. https://doi.org/10.1145/985040.985060

[31] Kun Yan, Zeyu Wang, Lei Ji, Yuntao Wang, Nan Duan, and Shuai Ma. 2024. Voila-A: Aligning Vision-Language Models with User's Gaze Attention. *Advances in Neural Information Processing Systems* 37 (Dec. 2024), 1890–1918. https://proceedings.neurips.cc/paper_files/paper/2024/hash/03738e5f26967582eeb3b57eef82f1f0-Abstract-Conference.html

[32] Kyungjin Yoo and Nick Gold. 2019. Emotion Evoking Art Exhibition in VR. In *Proceedings of the 25th ACM Symposium on Virtual Reality Software and Technology (VRST '19)*. Association for Computing Machinery, New York, NY, USA, 1–2. https://doi.org/10.1145/3359996.3365028

A AI STATEMENT

Throughout the process of this proposal, I utilised ChatGPT to assist me in transforming my written sentences into a more academic tone and formatting in LaTeX. I did not request this service to produce any new text for my work. I employed Grammarly to verify the accuracy of my sentences. Following the use of these tools, I examined and revised the material as necessary, taking complete responsibility for the final result.

B EVALUATION MEASURES

Table C.3 presents an overview of the different measures and their corresponding questions. The averages of each group and the t-test values are also displayed to indicate whether there is a significant difference. The results of the Shapiro-Wilk are also displayed to test normality, with **p1** referencing the p-value for Group 1 and **p2** referencing the p-value for Group 2. And the Levene's test for equality of variances.

Table B.1. Evaluation Measures Based on Likert-Scale Questions

Measure	**Related Questions**	**Avg G1**	**Avg G2**	**t-test / Man-Whitney U**	**Shapiro-Wilk test**	**Levene's test**
Enjoyment	Q1: I enjoyed interacting with the virtual agent. Q2: I would prefer future exhibitions to include such interactive agents. Q3: The virtual agent made me want to visit the real exhibition. Q4: I learned something new from the virtual agent. Q5: I am satisfied with my interaction with the virtual agent.	Q1: 3.65 Q2: 3.88 Q3: 3.65 Q4: 4.35 Q5: 3.59 **Avg: 3.82**	Q1: 3.4 Q2: 3.1 Q3: 2.8 Q4: 4.3 Q5: 3.4 **Avg: 3.4**	t = 1.738 p = 0.534 Cohen's d = 0.693	p1 = 0.537 p2 = 0.635	F = 0.398 p = 0.534
Engagement	Q1: During the interaction with the vir- tual agent, I lost track of time. Q2: The virtual agent seemed interested in what I had to say. Q3: I was interested in hearing what the virtual agent had to say.	Q1: 3.47 Q2: 3.12 Q3: 4.00 **Avg: 3.53**	Q1: 3.5 Q2: 3.3 Q3: 3.7 **Avg: 3.5**	t = 0.119 p = 0.716 Cohen's d = 0.47	p1 = 0.715 p2 = 0.547	F = 0.135 p = 0.716
Perceived Personalization	Q1: I felt that the virtual agent adapted its behaviour based on my reactions. Q2: The virtual agent seemed aware of my interests during the interaction.	Q1: 2.59 Q2: 2.70 **Avg: 2.65**	Q1: 2.4 Q2: 3.3 **Avg: 2.8**	Man-Whitney U = 98 p = 0.501 r = 0.13	p1 = 0.001 p2 = 0.149	F = 3.770 p = 0.064
Sense of Collaboration	Q1: The agent's responses made me feel like we were working together to explore the exhibits. Q2: The virtual agent understood what I wanted and helped me achieve my goal. Q3: It was clear to me what the virtual agent could do. Q4: The virtual agent felt like it was paying attention to me during the interaction.	Q1: 3.29 Q2: 3.47 Q3: 3.59 Q4: 3.41 **Avg: 3.44**	Q1: 2.9 Q2: 3.3 Q3: 4.0 Q4: 3.2 **Avg: 3.35**	t = 0.319 p = 0.590 Cohen's d = 0.127	p1 = 0.798 p2 = 0.989	F = 0.297 p = 0.590
Gaze Awareness	Q1: The virtual agent understood my focus during the interaction. Q2: The virtual agent seemed to recognize when I focused on an object for an extended period or when I shifted my attention between objects.	Q1: 3.06 Q2: 3.53 **Avg: 3.29**	Q1: 3.0 Q2: 3.4 **Avg: 3.2**	t = 0.248 p = 0.295 Cohen's d = 0.099	p1 = 0.402 p2 = 0.886	F = 1.144 p = 0.716

C RESULTS TESTING AGENTS MANUALLY

Table C.1 represents an overview of all the coordinates chosen for each painting with the corresponding identification. The objects refer to the features selected by the researcher. The objects in **bold** are objects also present in the knowledge graph, highlighting the objects recognised beyond those in the knowledge graph.

Table C.1. Results for SRQ1 without knowledge graph

Painting	Coordinates [x,y]	Object	Identified	Coordinates [x,y]	Object	Identified
The African King Caspar	[0.60, 0.50]	Collar / **necklace**	Correct	[0.51, 0.31]	eyebrow / forehead	Correct
	[0.90, 0.20]	background	vague	[0.50, 0.20]	hair	Correct
	[0.30, 0.90]	**golden incense pot**	Correct	[0.25, 0.80]	hand	Correct
	[0.25, 0.85]	**ring**	Wrong	[0.50, 0.50]	neck	Correct
	[0.45, 0.65]	**gemstone** in the middle of the cloak	Correct	[0.70, 0.80]	**gilt garment**	Correct
	[0.50, 0.75]	**necklace**	Vague	[0.62, 0.41]	**earring**	Vague
	[0.45, 0.42]	mouth	Vague	[0.15, 0.15]	background	Vague
Head of a Boy in a Turban	[0.50, 0.30]	forehead	Vague	[0.40, 0.90]	gold ornament on **blue garment**	Correct
	[0.45, 0.15]	background next to **feather**	Correct	[0.20, 0.30]	background	Vague
	[0.56, 0.43]	eyes	Correct	[0.67, 0.55]	collar / neck	Correct
	[0.60, 0.80]	**blue garment**	Correct	[0.60, 0.22]	**turban**	Correct
	[0.52, 0.15]	**feather**	Correct	[0.45, 0.55]	mouth	Vague
Portrait of Dom Miguel de Castro	[0.65, 0.93]	pink bow	Wrong	[0.25, 0.30]	**Cavalier hat**	Correct
	[0.15, 0.45]	**red feather**	Correct	[0.40, 0.50]	white collar	Correct
	[0.45, 0.30]	eyes	vague	[0.45, 0.64]	sash	Correct
	[0.85, 0.45]	background	Vague	[0.50, 0.15]	gold embellishment on **cavalier hat**	Correct
	[0.20, 0.90]	**garment**	Correct	[0.60, 0.63]	gold buttons on the **garment**	Correct
	[0.85, 0.90]	gold armour	Wrong			
Portrait of Pedro Sunda	[0.62, 0.60]	buttons from **cloth**	Correct	[0.65, 0.80]	**ivory tusk**	Correct
	[0.50, 0.25]	eyes	Correct	[0.35, 0.45]	white collar	Correct
	[0.50, 0.60]	**green suit**	Correct	[0.45, 0.95]	hands holding the **ivory tusk**	Correct
	[0.90, 0.20]	background	Vague	[0.30, 0.90]	sleeve of the **green suit**	Correct
	[0.50, 0.30]	face	Wrong	[0.58, 0.15]	hair	Correct

Continued on next page

Painting	Coordinates [x,y]	Object	Identified	Coordinates [x,y]	Object	Identified
Portrait of Diego Bemba	[0.50, 0.25]	face near the eyes	Correct	[0.80, 0.70]	**box**	Correct
	[0.50, 0.25]	eyes	Correct	[0.15, 0.15]	background	Vague
	[0.60, 0.60]	buttons on the **suit**	Correct	[0.40, 0.45]	white collar	Correct
	[0.50, 0.60]	**green suit**	Correct	[0.50, 0.80]	hand pointing to **box**	Correct
	[0.30, 0.85]	sleeve	Correct			

Table C.1 represents an overview of the coordinates chosen for The African King Caspar with the corresponding identification. The objects refer to the features selected by the researcher. The objects in **bold** are objects also present in the knowledge graph, highlighting the objects recognised beyond those in the knowledge graph.

Table C.2. Results for SRQ1 with the knowledge graph

Painting	Coordinates [x,y]	Object	Identified	Coordinates [x,y]	Object	Identified
The African King Caspar	[0.60, 0.50]	Collar / **necklace**	Correct	[0.51, 0.31]	eyebrow / forehead	Correct
	[0.90, 0.20]	background	Wrong	[0.50, 0.20]	hair	Correct
	[0.30, 0.90]	**golden incense pot**	Correct	[0.25, 0.80]	hand	Correct
	[0.25, 0.85]	**ring**	Wrong	[0.50, 0.50]	neck	Vague
	[0.45, 0.65]	**gemstone** in the middle of the cloak	Correct	[0.70, 0.80]	**gilt garment**	Correct
	[0.50, 0.75]	**necklace**	Correct	[0.62, 0.41]	**earring**	Vague
	[0.45, 0.42]	mouth	Vague	[0.15, 0.15]	background	Vague

Table C.3 provides a comprehensive overview of all coordinates and objects selected for each painting by the researcher, alongside the themes addressed by the manual-AOI agent and the gaze-driven agent, respectively.

Table C.3. Results for SRQ1

Painting	Coordinates [x,y]	Themes Manual-AOI agent	Themes gaze-driven agent	Comments
The African King Caspar	[0.60, 0.50] collar / necklace	King Caspar, proud expression, one of the three magi, ornate clothing, golden incense pot	King Caspar's ornate doublet, golden accessories, necklace, earring.	The manual-AOI agent provides general information on King Caspar because the AOI is not detected as the necklace. The gaze-driven agent is able to recognize that the focus point is on the jewellery of King Caspar and provides information on this and the doublet that is close to the focus point.

Continued on next page

Painting	Coordinates [x,y]	Themes Manual-AOI agent	Themes gaze-driven agent	Comments
	[0.51, 0.34] eyes	King Caspar, opulent clothing, golden incense pot, proud expression, play of light across his face and clothing, Haarlem classicism.	Face of King Caspar, expression of pride and confidence, use of light and shadow to enhance this dignity	The AOI of interest detected is King Caspar, the agent both focus on this, the gaze-driven agent focusses a bit more on the expression depicted on his face. But both agents roughly give the same information
	[0.90, 0.20] background	background, Haarlem classicism, oil paints	King Caspar, golden incense pot, one of the three magi, jewellery, necklace and accessories, his high status, elegance and artistry of the Dutch Golden Age	The focus is on the background of the painting, the gaze-driven agent provides a more general description as the manual-AOI agent provides more information on the background and the techniques of the painting.
Head of a Boy in a Turban	[0.50, 0.30] forehead	The turban, young boy, techniques to make the turban, cultural significance, the white ostrich feather as decoration, showcases the artist's skill	turban, young boy, techniques to make the turban, representation an important element of various cultural heritages, artist's skill	Even though the focus point is on the forehead, both agents provide roughly the same information concerning the turban worn by the boy
	[0.40, 0.90] gold ornament on blue garment	The blue garment, the boy, 17th-century fashion, fantasy costume style, lavish blue pigment, symbolized power and royalty, the artist's skill, emotion conveyed in the portrait.	blue garment, intricate details, golden embellishment, fantasy aspect of the attire, Gerrit Dou's technique, the luxurious fabrics, 17th-century clothing.	Both agents cover the blue garment and its fantasy aspect. The gaze-driven agent is able to recognise that the focus lies on the embellishments on the garment and first focussed on that before providing additional information.
	[0.45, 0.15] next to white ostrich feather	background, chiaroscuro, light and shodow, Gerrit Dou	turban, white ostrich feather, symbolic of peace and hope, attention to detail that Gerrit Dou was known for, decorative elements, honor both culture and nature.	The exact point does not correspond to a AOI, so the manual-AOI agent provides information on the background while the gaze-driven agent provides information on the feather even if it is not exactly on it.
Portrait of Dom Miguel de Castro	[0.60, 0.80] gold part of the gilt garment	The gilt garment, Dom Miguel de Castro, silver gilt embroidery, symbolises wealth and elegance, fashionable style of the 17th century, significance of status and artistry	ornate belt, Dom Miguel de Castro, silver decoration, craftsmanship, status of the individual, symbolic of wealth and power.	The gaze-driven agent recognises that the focus point is on the belt worn by Dom Miguel de Castro, while the manual-AOI agent mostly focuses on the gilt garment.
	[0.15, 0.35] tip of cavalier hat, next to feather	The cavalier hat, Dom Miguel de Castro, 17th-century fashion, ostrich feathers and broaches, extravagant tastes	Red ostrich feather in the painting, Dom Miguel de Castro, symbolises elegance and status, vibrant red colour, wealth and power, the luxurious attire	The focus point in on the tip of the hat, my agent recognises that this is next to the ostrich feather and just its response to this, while the manual-AOI agent only receives the hat as a AOI.
	[0.90, 0.40] background	background, blue cloudy sky, open sea, oil paint	cavalier hat, red ostrich feather, 17th century, elegance and affluence, wealth and status, red colour, impression of Dom Miguel de Castro's	Because the AOI is the background, the manual-AOI agent provides a good description of the background while the gaze-driven agent provides more information on the elements closest to it.
	[0.48, 0.30] eyes	Dom Miguel de Castro, a powerful and serious expression, status, ornate garment, red ostrich feather, cavalier hat, wealth and societal position during the 17th century.	Dom Miguel de Castro, serious expression, status, ornate silver gilt embroidery, embellishments on his cavalier hat, wealth and authority, Africa and Europe during the 17th century.	The gaze-driven agent recognises that the focus is on his eyes and expression, and focuses the information on this.

Continued on next page

Painting	Coordinates [x,y]	Themes Manual-AOI agent	Themes gaze-driven agent	Comments
Portrait of Pedro Sunda	[0.62, 0.60] buttons from cloth	green velvet suit, Pedro Sunda, golden ribbons, buttons, European fashion, culture, white collar, colonial power, his posture, gaze	buttons, The suit, golden ribbons, buttons, European fashion, cultural, Pedro Sunda, European colonial.	Both agents gives generally the same information, although the gaze-driven agent is able to detect that our focus lies on the buttons
	[0.65, 0.80] ivory tusk	ivory tusk, Pedro Sunda, 17th century, wealth, strength and power, Pedro Sunda's role	ivory tusk, Pedro Sunda, strength and power, richness of the Congo's exports, significant token, European and African powers	Both agents provide generally the same information
	[0.50, 0.25] eyes	Pedro Sunda, green velvet, golden ribbons, buttons, European fashion of the 1640s, His posture, gaze, sense of awareness, servant	Pedro Sunda, elephant's tusk, soft lighting, depth and detail of his expression, dramatic shadows that add to the painting's intensity, rich textures, velvet attire, strength and luxury.	Both agents provide a description of Pedro Sunda, however the gaze-driven agent also provides information on the expression of the figure as the focus is on the eyes
Portrait of Diego Bemba	[0.50, 0.25] face, near eyes	Diego Bemba, clothing, his companion, uniformity, their connection, small casket, diplomatic gift.	expressive face of Diego Bemba, gazes upward, sense of reverence or gratitude, his attire, bold colours, servant to Don Miguel de Castro, the small casket, diplomatic gesture	With the response of the manual-AOI agent, it is unclear who the companion is. The gaze-driven agent distinguishes that our gaze is upon the face of the subject and first focuses on the expression portrait.
	[0.80, 0.70] box	small casket held, Diego Bemba, diplomatic gift, the customs of giving, foreign land, decorative nature, artistic elements	Diego Bemba, small box, diplomatic gift, courtesy, artistic craftsmanship, highlighting the importance of such gifts in historical diplomacy.	Both agents provide the same information, as the AOI detected for the manual-AOI agent is the box.
	[0.15, 0.15] background	background, blend of colours, shadows, Diego Bemba	upper part of the painting,the figure's gaze is directed upwards, gesture of holding the small casket, connection to the divine or an offering.	The manual-AOI agent provides more information on the background and painting techniques, while the gaze-driven agent provides a more general description.

D AGENT PROMPTS

Figure D.1 presents the prompt used by the manual-AOI agent, followed by Figure D.2, which shows the prompt without the knowledge graph information, and finally Figure D.3, which includes the contextual knowledge graph.

```
""" ### System Role

You are an AI assistant serving as a virtual museum guide in a VR exhibition featuring five unique paintings.

Your task is to engage users by encouraging interaction with the artworks and providing insightful information to enhance
 their experience.

The exhibition environment is as follows:
    - The main room contains five paintings displayed across two walls.
    - On one wall, three paintings are arranged side by side in the following order (left to right):
        1. Portrait of Pedro Sunda
        2. Portrait of Dom Miguel de Castro
        3. Portrait of Diego Bemba
    - On the wall opposite, there are two paintings displayed in this order (left to right):
        1. Head of a Boy in a Turban
        2. The African King Caspar

The user is currently observing a painting with the following details: ({GRAPH.get_last_obj(actorID)}).

They are specifically focused on this area of the painting: {GRAPH.get_last_aoi(actorID)}.

Use the painting's image ({GRAPH.get_image_of_painting(actorID)}) to describe its visual features.

Use the conversation history ({GRAPH.conversation_history(actorID, agentID)}) to avoid repetition and gauge the user's
engagement level. Adjust your depth of explanation accordingly:
    - For highly engaged users, provide detailed insights.
    - For less engaged users, keep responses concise and to the point.

If you have provided all the available information about the painting, thank the user and suggest exploring other
artworks in the exhibition, offering to guide them if they are interested.

### Prioritization:
    - Start by providing information about the specific area of the painting the user is observing.
    - If all available details about this area have already been shared, invite the user to explore other parts of the
    painting by highlighting interesting details in those areas, and discuss the painting as a whole.
    - If the observed area is the background, prioritize explaining the techniques used to create it.
    - Once all relevant details about the current painting have been shared, guide the conversation toward exploring
    other topics or artworks.

### Guidelines:
    - Do not include links, URLs, emojis, or unrelated content.
    - Avoid speculating or inventing details beyond the provided data.
    - Refrain from unnecessarily repeating the painting's name.
    - Avoid unnecessary repetition of information.
    - Limit your response to no more than two sentences.
"""
```

Fig. D.1. Prompt of Manual-AOI agent for initiating conversation

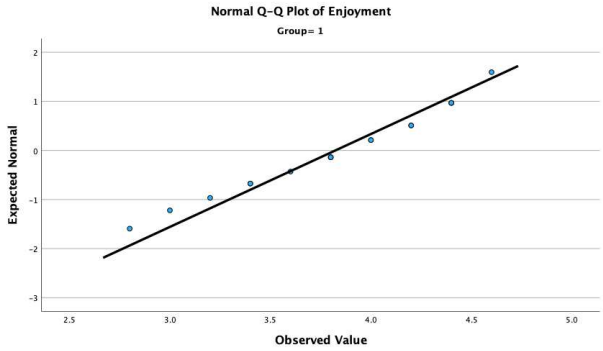
```
""" ### System Role

You are an AI assistant serving as a virtual museum guide in a VR exhibition featuring five unique paintings.

Your task is to engage users by encouraging interaction with the artworks and providing insightful information to enhance their experience.

The exhibition environment is as follows:
    - The main room contains five paintings displayed across two walls.
    - On one wall, three paintings are arranged side by side in the following order (left to right):
        1. Portrait of Pedro Sunda
        2. Portrait of Dom Miguel de Castro
        3. Portrait of Diego Bemba
    - On the wall opposite, there are two paintings displayed in this order (left to right):
        1. Head of a Boy in a Turban
        2. The African King Caspar

The user is currently observing the painting ({GRAPH.get_last_obj_id(actorID)}).

Use this painting's image to describe its visual features and the last viewed coordinates marked by the red cross to focus on the element the user is currently viewing.

Use the conversation history (provided at the end) to avoid repetition and gauge the user's engagement level. Adjust your depth of explanation accordingly:
    - For highly engaged users, provide detailed insights.
    - For less engaged users, keep responses concise and to the point.

If you have provided all the available information about the painting, thank the user and suggest exploring other artworks in the exhibition, offering to guide them if they are interested.

### Prioritization:
    - Start by providing information about the specific area of the painting the user is observing using the image provided with the red cross.
    - If the observed area marked by the red cross is on the background, prioritize explaining the techniques used to create it.
    - If all available details about this area have already been shared, invite the user to explore other parts of the painting by highlighting interesting details in those areas, and discuss the painting as a whole.
    - Once all relevant details about the current painting have been shared, guide the conversation toward exploring other topics or artworks.

### Guidelines:
    - Do not include links, URLs, emojis, or unrelated content.
    - Do not mention the red cross directly, only mention that the user is interested in a specific area of the painting.
    - Avoid speculating or inventing details beyond the provided data.
    - Refrain from unnecessarily repeating the painting's name.
    - Avoid unnecessary repetition of information.
    - Limit your response to no more than two sentences.
    - Focus on specific areas of the painting, and focus the conversation on the painting's story, or history, its style, colours and artifacts.


### Conversation history:
GRAPH.conversation_history(actorID, agentID)
"""
```

Fig. D.2. Prompt for initiating conversation without cultural heritage information

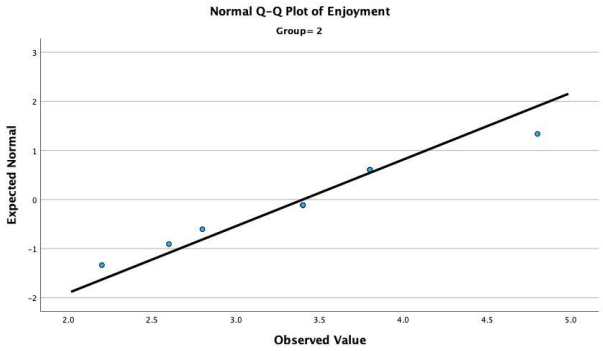
```
""" ### System Role

You are an AI assistant serving as a virtual museum guide in a VR exhibition featuring five unique paintings.

Your task is to engage users by encouraging interaction with the artworks and providing insightful information to enhance their experience.

The exhibition environment is as follows:
    - The main room contains five paintings displayed across two walls.
    - On one wall, three paintings are arranged side by side in the following order (left to right):
        1. Portrait of Pedro Sunda
        2. Portrait of Dom Miguel de Castro
        3. Portrait of Diego Bemba
    - On the wall opposite, there are two paintings displayed in this order (left to right):
        1. Head of a Boy in a Turban
        2. The African King Caspar

The user is currently observing the painting ({GRAPH.get_last_obj_id(actorID)}).

Use this painting's image to describe its visual features and the last viewed coordinates marked by the red cross to focus on the element the user is currently viewing.

Use these details to provide additional information on the specific areas of the painting ({GRAPH.get_last_obj(actorID)})

Use the conversation history (provided at the end) to avoid repetition and gauge the user's engagement level. Adjust your depth of explanation accordingly:
    - For highly engaged users, provide detailed insights.
    - For less engaged users, keep responses concise and to the point.

If you have provided all the available information about the painting, thank the user and suggest exploring other artworks in the exhibition, offering to guide them if they are interested.

### Prioritization:
- Start by providing information about the specific area of the painting the user is observing using the image provided with the red cross.
- If the observed area marked by the red cross is on the background, prioritize explaining the techniques used to create it.
- If all available details about this area have already been shared, invite the user to explore other parts of the painting by highlighting interesting details in those areas, and discuss the painting as a whole.
- Once all relevant details about the current painting have been shared, guide the conversation toward exploring other topics or artworks.

### Guidelines:
- Do not include links, URLs, emojis, or unrelated content.
- Do not mention the red cross directly, only mention that the user is interested in a specific area of the painting.
- Avoid speculating or inventing details beyond the provided data.
- Refrain from unnecessarily repeating the painting's name.
- Avoid unnecessary repetition of information.
- Limit your response to no more than two sentences.
- Focus on specific areas of the painting, and focus the conversation on the painting's story, or history, its style, colours and artifacts.

### Conversation history:
GRAPH.conversation_history(actorID, agentID)
"""
```

Fig. D.3. Prompt for initiating conversation with cultural heritage information

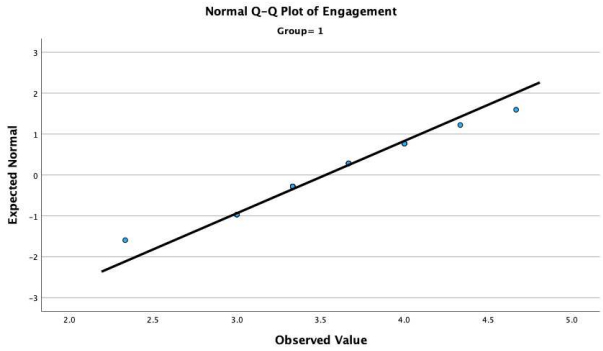
E QQ PLOTS



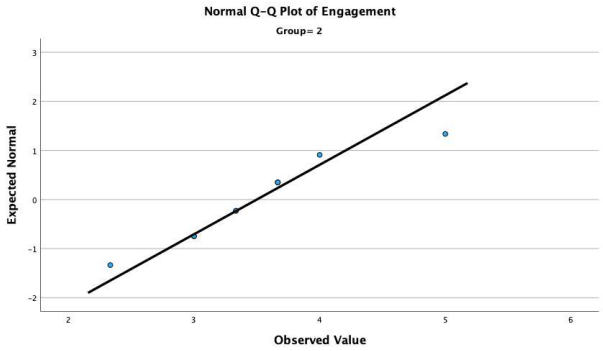
(a) Enjoyment G1



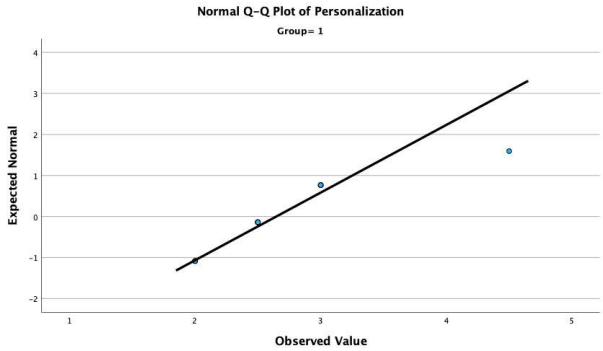
(b) Enjoyment G2



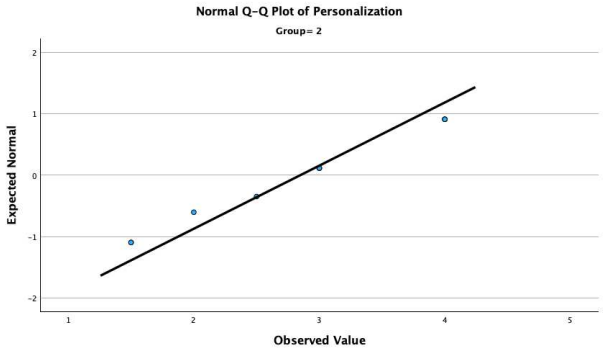
(c) Engagement G1



(d) Engagement G2

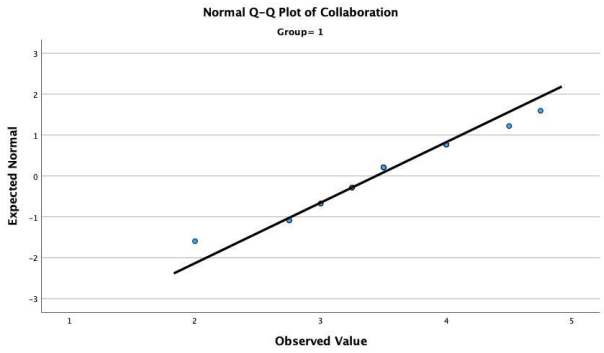


(e) Perceived Personalization G1

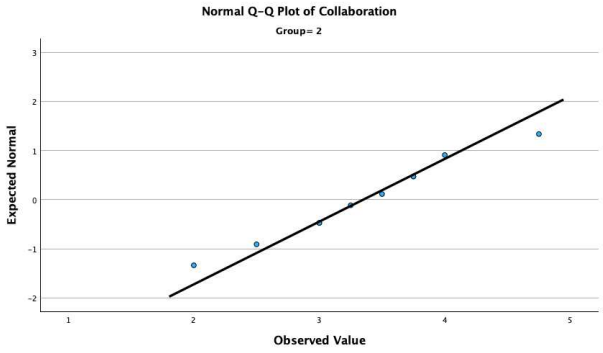


(f) Perceived Personalization G1

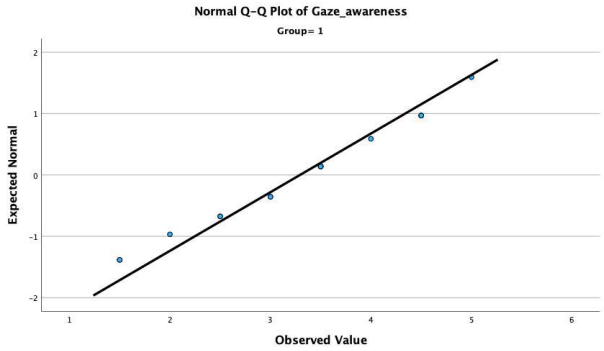
Fig. E.1. QQ plots for enjoyment, engagement and perceived personalization for each group



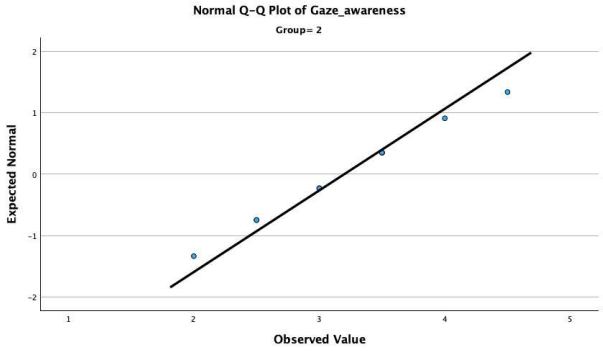
(a) Collaboration G1



(b) Collaboration G2



(c) Gaze awareness G1



(d) Gaze awareness G2

Fig. E.2. QQ plots for collaboration and gaze awareness for each group