Multimodal Emotion Elicitation and Recognition in Virtual Reality

ZHEYUAN KUANG

Supervisor: Dr. Zhanna Sarsenbayeva
Associate Supervisor: Prof. Anusha Withana

A thesis submitted in fulfilment of
the requirements for the degree of
Master of Philosophy

School of Computer Science
Faculty of Engineering
The University of Sydney
Australia

21 May 2026

Statement of Originality

This is to certify that the content of this thesis is my own work. This thesis has not been submitted for any other degree or purpose.

I, Zheyuan Kuang, certify that the intellectual content of this thesis is the product of my own work, and that all assistance received in preparing this thesis and all sources have been acknowledged.

Name: Zheyuan Kuang

Date: 21 May 2026

Abstract

Virtual Reality (VR) has been effectively used for eliciting emotions, yet most research focuses on the intensity of affective responses rather than on how interaction influences those experiences. To address this gap, this thesis advances a validated VR emotion-elicitation dataset through two extensions. First, we add a new high-arousal, high-valence scene and validate its effectiveness in a within-subject study (N=24). Second, we incorporate interactive elements into each scene by creating interactive and non-interactive versions to examine the impact of interaction on emotional responses. We evaluate interaction through a multimodal approach combining subjective ratings and physiological signals. Our evaluation study (N=84) shows that interaction not only amplifies emotions but also modulates them in context, supporting coping in negative scenes and enhancing enjoyment in positive scenes.

Multimodal Emotion Recognition (MER) increasingly depends on fine-grained, evidence-grounded annotations, yet inspection and label construction are hard to scale when cues are dynamic and misaligned across modalities. This thesis presents an LLM-assisted toolkit that supports multimodal emotion data annotation through an inspectable, event-centered workflow. The toolkit aligns heterogeneous recordings, visualizes modalities on an interactive shared timeline, and packages synchronized keyframes and time windows as traceable event packets. It then integrates an LLM with modality-specific tools and prompt templates to draft structured annotations for analyst verification and editing.

Building on the dataset extensions and the event-centered annotation outputs, this thesis further investigates modeling approaches for MER in VR that integrate behavioural and physiological signals collected from VR headsets and wearable sensors. In particular, we introduce an LLM-based Mixture-of-Experts (MoE) framework for multimodal emotion recognition, where experts specialize in different modalities and a router assigns weights to experts for each event. The goal is to better connect predictions to traceable multimodal evidence and support interpretation of affective cues in interactive VR settings.

Acknowledgements

First, I would like to express my deepest gratitude to my supervisor, Dr. Zhanna Sarsenbayeva, for her steady guidance, unwavering support, and constant encouragement throughout my MPhil. Her help has been fundamental to this achievement. She has taught me not only how to do research, but also how to be a kind person and a hardworking student. She continues to remind me how to think clearly, work responsibly, and manage time with purpose. The way she meets the demands of research with both rigour and flexibility has shaped my view of what good research can be, and I am sincerely grateful.

I would also like to extend my heartfelt thanks to Prof. Anusha Withana, Prof. Weiwei Jiang, Dr. Benjamin Tag, Prof. Flora Salim, and Prof. Sven Mayer. Their enthusiasm, generosity, and thoughtful conversations brought clarity to my thinking. Each discussion, question, and suggestion helped me move forward, often turning uncertainty into direction and making the research journey brighter day by day.

Special thanks go to my friends and colleagues: Tinghui, Yihao, Wenqi, Rocky, Adele, Marvin, Frank, Lefan, Yixiang, Hanyang, Shengzhe, Wendi, Shakyani, Praneeth, Pamuditha, and all AID Lab members. Thank you for your support, your patience, and your words of encouragement throughout my MPhil. Your presence, humour, and kindness carried me through long days, and your companionship made this journey feel shared rather than solitary.

Lastly, and most importantly, I am grateful to my family, whose love, care, patience, and belief in me have been a constant source of strength. They made the difficult days lighter and the good days more meaningful. Without them, this work would not have been possible. Thank you for walking beside me through every step, and for navigating life with me, now and always.

Author Attribution Statement

This thesis is based on publications and works in progress, for which I am the lead author and have made key contributions to each. Additionally, all studies presented in this thesis received ethics approval under project number 2025/HE00028.

- **Zheyuan Kuang**, Tinghui Li, Weiwei Jiang, Sven Mayer, Flora D. Salim, Benjamin Tag, Anusha Withana, and Zhanna Sarsenbayeva. 2026. Understanding the Effects of Interaction on Emotional Experiences in VR. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems.* CHI '26, Barcelona, Spain. https://doi.org/10.1145/3772318.3790313

  This paper is accepted for publication in CHI'26 conference. My contributions included proposing the concept, developing the scenes, conducting the simulations and experiments, and writing the drafts.

- **Zheyuan Kuang**, Weiwei Jiang, Nicholas Koemel, Matthew Ahmadi, Emmanuel Stamatakis, Benjamin Tag, Anusha Withana, and Zhanna Sarsenbayeva. 2026. An LLM-Assisted Toolkit for Inspectable Multimodal Emotion Data Annotation. Accepted to the CHI Workshop on *Co-Data: Cultivating Effective Human-LLM Collaboration for Collaborative Data Processing*. CHI '26, Barcelona, Spain.

  This paper is accepted to the CHI'26 workshop. My contributions included proposing the concept, designing the toolkit, and writing the drafts.

Zheyuan Kuang

Dr. Zhanna Sarsenbayeva

21 May 2026

Use of Generative AI Statement

During the preparation of this thesis, I used ChatGPT (OpenAI, `https://chatgpt.com/`) for limited editorial assistance, including grammar correction, improvement of sentence structure, clarity, and academic wording. All AI-assisted outputs were reviewed critically for possible errors, inaccuracies, and bias, and revised as necessary. I take full responsibility for the final content of this thesis.

CONTENTS

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Background and Motivation

Emotions play a central role in shaping human experience, influencing perception, decision-making, and behaviour [63]. Virtual reality (VR) has emerged as a powerful tool for eliciting emotions, offering immersive and ecologically valid environments where affective states can be readily elicited for integrated emotion experiences [46]. Prior research has successfully used immersive 360° videos, photorealistic environments, and virtual environments to evoke a range of emotions [65, 108, 46]. However, the majority of existing methods rely on passive stimuli such as visual and audio, with limited exploration of interactive or active forms of emotion elicitation in VR [8]. Compared with passive exposure, interactivity introduces users' actions, feedback, and sense of agency into the emotion elicitation process, allowing emotional responses to be shaped not only by external stimuli, but also by human's actions and engagement with the environment [63, 102]. As a natural extension of VR, interactive VR environments hold promise for enhancing emotional experiences.

While prior work using virtual environments supports scene-level interaction (*e.g.*, teleportation), the specific role of object-level interaction in modulating emotional responses remains underexplored [20, 113, 46]. In particular, research on emotion in VR has focused on measuring responses and designing affective visualizations, yet few studies have systematically varied the interaction to modulate and potentially enhance its effect on emotional engagement. However, previous studies suggest that object-level interaction holds significant potential. For example, object-level interaction enables VR users to reach out and touch objects, making them much more immersed in the game world than traditional screens [55, 49]. The opportunity to touch an object increases the feeling of perceived ownership of that object, and the valuation of the object is also increased when the touch experience provides either neutral or positive sensory feedback [98]. Novel designs for object interaction can significantly enhance

perceived fun and user satisfaction [135]. Moreover, objects in the virtual world can exert variable impact or control on users' decisions, and thus object-level interaction has an impact on dominance by suggesting or preventing users from accomplishing actions, as does a lack of interaction with the environment [26]. Despite these insights, it remains unclear how these mechanisms influence emotional experience in VR. Investigating this relationship can clarify the affective role of object-level interaction and guide the design of more emotionally engaging VR environments, which could be applied in mental health research and practice [114, 9]. In this thesis, we first explore the fundamental question: *Does enabling object-level interaction in virtual environments enhance the elicitation of emotional experiences?*

To investigate this, we extend the VR emotion-elicitation dataset by Jiang et al. [46] in two ways. First, we integrate affectively-designed interactive elements tailored to each scene by Jiang et al. [46]. For our experimental comparison, each scene is implemented in two versions: an interactive version featuring user-triggered actions (*e.g.*, petting puppies, throwing stones, using a shield) and a non-interactive version with identical audiovisual content. Furthermore, we extend the existing dataset [46] by adding an additional scene. The authors reported that the dataset lacked a scene in the High-Arousal High-Valence (HAHV) quadrant of Russell's circumplex model [104], which limited its coverage of the full affective space. To address this limitation, we developed a new scene, *Surrounded by Elephants*, based on a validated 360° video [65]. This addition ensures that all four quadrants are now represented, enabling more balanced emotion elicitation. We have implemented both interactive and non-interactive versions of the scene to investigate the effect of object-level interaction on emotion elicitation. We collect multimodal recordings during these VR experiences. To support downstream analysis and modeling, this thesis investigates LLM-assisted multimodal annotation and LLM-based MER. Together, these components form a pipeline from controlled VR emotion elicitation to inspectable, event-centered annotation and evidence-based multimodal emotion recognition.

Working with these multimodal VR recordings is often constrained by data work. Emotion-related cues are distributed across modalities and vary over time, making cross-modal inspection, integration, and event-level interpretation difficult. This increases the demand for fine-grained, descriptive, and evidence-grounded annotations, but benchmark results highlight that annotation cost, limited labelled data, and reduced robustness under domain gaps, noise, and missing modalities remain major barriers [73, 75]. Motivated by these needs, this thesis develops an LLM-assisted toolkit that supports inspectable, event-centered multimodal emotion data annotation. The toolkit assists analysts in aligning heterogeneous

recordings, selecting candidate events, and drafting structured annotations with traceable links to source evidence, while keeping humans in control for verification and editing [143, 122].

These foundations also motivate modeling work for multimodal emotion recognition (MER). Conventional deep learning models for behavioural and physiological signals often require task-specific training and can generalize poorly across contexts, while their outputs are typically hard labels without transparent reasoning [133, 132, 86, 144]. Recent studies indicate that LLMs can support multimodal reasoning and explanation across video and sensor data [144, 59, 109, 19, 70, 137], but LLM-based MER remains underexplored in immersive VR settings. To address this gap, this thesis investigates an LLM-based MER framework that integrates behavioural and physiological signals collected during VR experiences. In particular, we incorporate a Mixture-of-Experts (MoE) module to support modality-adaptive multimodal fusion and to provide human-understandable explanations for MER.

Across our studies, we evaluate emotional responses using both subjective self-reports and objective physiological measures. While the Self-Assessment Manikin (SAM) questionnaire provides valuable subjective ratings of valence and arousal, it has inherent limitations [14]. It is a discrete, post-hoc measure that is susceptible to cognitive biases and recall inaccuracies, and it may fail to capture the full intensity of transient, high-arousal states. Physiological sensing, in contrast, provides an objective, continuous, and unconscious measure of affective responses, particularly for arousal-related responses [106]. This is especially critical for our study, as the visceral, immediate impact of interactive elements – like the surprise of an elephant's reaction or the tension of raising a shield – may manifest in the autonomic nervous system before they are consciously processed and reported. By integrating physiological data, we aim to capture these nuanced, real-time reactions, providing a more complete and robust understanding of how interaction modulates emotional experience.

This work aims to explore the following research questions:

RQ1 Does the added scene *Surrounded by Elephants* reliably and effectively elicit High Arousal and High Valence emotions?

RQ2 How does object-level interaction influence subjective and physiological measures of emotional response compared to a non-interactive baseline?

RQ3 What is the relationship between subjective self-reports and physiological arousal in response to affective interactions in VR?

RQ4 How can an LLM-assisted toolkit support inspectable and event-centered multimodal emotion data annotation?

RQ5 How can an LLM-based MoE framework improve multimodal fusion and provide human-understandable explanations for MER?

1.2 Main Contribution

In summary, our work makes five key contributions to the fields of HCI and affective computing:

- We provide an extended VR dataset[1] with a new scene that fills a critical gap in eliciting high-arousal, high-valence emotions, enabling richer and more balanced affective research.
- We provide one of the first systematic investigations into the effects of object-level interaction on emotional response, using a controlled experimental design with interactive and non-interactive versions of each scene.
- We provide an empirical analysis combining self-report and physiological data, offering nuanced insights into affective processes and propose practical guidance for designing emotionally resonant interactive VR experiences.
- We introduce an LLM-assisted toolkit[2] that supports inspectable, event-centered multimodal emotion data annotation, enabling scalable inspection and traceable label construction through human verification.
- We propose an LLM-based MoE framework that aims to improve multimodal fusion and provides human-understandable, evidence-linked explanations for multimodal emotion recognition.

1.3 Thesis Outline

More specific contributions from each chapter are organized as follows.

Chapter 2 reviews prior work on measuring emotions, eliciting emotions in VR, affective interaction, emotion recognition in virtual reality, LLM-assisted data annotation, using LLMs for emotion recognition, and Mixture-of-Experts.

[1] https://github.com/ZHEYUANK/VR-Dataset-Emotions-Interaction.git
[2] https://github.com/ZHEYUANK/Multimodal-Reviewer.git

Chapter 3 presents our empirical work on interaction and emotional experiences in VR. It first describes the design and modeling of the VR scenes, then reports Study 1 for validating the effectiveness of the *Surrounded by Elephants* scene, and Study 2 for examining how object-level interaction influences emotion elicitation. The chapter concludes with evaluation results covering emotion elicitation measures, engagement time, engagement with the virtual environments, physiological measures, and topic modeling of participants' responses.

Chapter 4 presents an LLM-assisted toolkit for inspectable multimodal emotion data annotation, including preprocessing and visualization, event detection and retrieval, and LLM-based annotation support.

Chapter 5 introduces a Mixture-of-Experts LLM framework for multimodal emotion recognition in virtual reality, and describes its methodology.

Chapter 6 discusses the thesis findings by addressing the research questions, positioning our results within existing literature, outlining implications for research and design, and noting limitations and future work.

Chapter 7 concludes the thesis by summarizing the main outcomes and contributions.

CHAPTER 2

Literature Review

In this chapter, we review key areas relevant to this thesis, including methods for measuring emotions, approaches to emotion elicitation in VR, affective interaction, multimodal emotion recognition, and LLM-based methods for annotation and recognition.

2.1 Measuring Emotions

The widely used Circumplex Model of Affect [104] conceptualizes emotions along two continuous dimensions: valence and arousal. Ekman's model [27] classifies emotions into discrete basic categories; however, its universality remains debated due to, e.g., cultural variations [11]. Building on these theoretical frameworks, emotions in HCI research are often measured subjectively using psychometric self-report scales. Common approaches include categorical emotion scales [24], the Positive and Negative Affect Schedule (PANAS) [124], the Self-Assessment Manikin (SAM) [14], the Pleasure-Arousal-Dominance (PAD) scale [89], and the Affect Slider [11]. PANAS combines valence and arousal into positive and negative affect scores but may misinterpret pleasure-driven positive affect, especially in high-arousal negative scenarios such as anger, leading to misleadingly high positive scores [22, 99, 125]. SAM introduces a pictorial assessment with pleasure (valence), arousal, and dominance dimensions, providing an intuitive and efficient tool for self-reporting emotions [14]. Recent studies have demonstrated its robustness and reliability in VR environment validations [130]. Given its multidimensional nature and image-based icons, SAM is particularly suitable for immersive VR contexts, and we adopt it in our study to capture valence, arousal, and dominance more comprehensively.

Objective emotion measures mostly rely on wearables and physiological sensors, as these provide convenient ways to estimate emotions by measuring multiple signals linked to the central and autonomic nervous systems [133], including electroencephalography (EEG) [115], heart rate (HR) and heart rate variability (HRV) [123], electrodermal activity (EDA) [6], facial behaviors [105, 134, 116], and eye

blink [142]. However, these signals can be difficult to interpret, as similar patterns may reflect different emotional states and are often influenced by noise and user movement, especially in immersive VR settings [4, 117]. Despite offering objective insights, physiological signals do not fully capture the subjective experience of emotion and require careful preprocessing and context-specific interpretation [100]. To improve reliability, combining multiple physiological signals can enhance emotion recognition accuracy and allow for richer measurement of affective states [38]. In addition, VR provides a highly controllable and repeatable setting where such multimodal measurements can be precisely integrated and validated [68, 69], supporting more comprehensive assessments of emotional responses [66, 128, 67, 101].

2.2 Eliciting Emotions in VR

Emotion elicitation in VR has traditionally relied on brief and isolated virtual environments designed to induce specific affective states, typically through brief exposures lasting only a few minutes [30, 47, 86]. These environments often incorporate visual stimuli such as images from the International Affective Picture System (IAPS) [60] or 360° videos [108], alongside auditory cues like emotionally expressive music [110]. Beyond perceptual stimuli, recent studies have adapted autobiographical recall into VR, using immersive cue-based paradigms to trigger emotional memories and associated physiological responses [37].

While effective for controlled induction, these approaches often lack interaction and dynamic affective progression. Recent studies have explored more immersive experiences, particularly VR games, which unfold over extended periods and evoke evolving emotional trajectories [35, 42]. These experiences span a range of valence and arousal levels [10, 92], and can be shaped by design factors such as avatar expressions [51], ambient lighting and color [7], and music [54]. However, the complexity and duration of VR games may introduce variability in emotional responses and reduce their suitability for tightly controlled experimental studies [62, 64].

Beyond basic emotional triggers, recent work has begun designing immersive VR environments to elicit more complex states such as awe [20] and has provided evidence that emotionally responsive experiences can be achieved through deliberate virtual environment design [113]. As a highly immersive, interactive, and customizable medium, VR offers a promising platform for studying how specific design factors shape emotional experience [46].

2.3 Affective Interaction

Affective interaction refers to the ways in which emotionally meaningful exchanges occur between a user and a designed system [79]. While many VR emotion studies rely on passive exposure, recent work has begun to explore how interactive elements shape emotional experience. Studies have shown that features such as haptic feedback [29], auditory signals [39], and interaction with virtual objects [76], virtual agents [44], virtual pets [95] can modulate affective responses. For example, throwing paper planes in VR can help users symbolically release negative emotions [120]. Such interactive designs have been shown to enhance emotional engagement by enabling natural and goal-directed actions in immersive VR settings [113]. Furthermore, researchers demonstrated that presence and emotion reinforce each other during interaction [102], further supporting the role of interaction in affective experience. However, most existing studies are designed for specific applications and lack reusable resources to systematically study how interaction design in VR contributes to emotional responses. Our work addresses this limitation by extending a validated VR emotion-elicitation dataset to cover all four affective quadrants and providing scene-tailored object-level interaction designs. It examines how affective interaction influences user engagement and emotional responses in VR-based emotion elicitation, while offering reusable materials and a baseline for future studies.

2.4 Multimodal Emotion Recognition

Multimodal Emotion Recognition (MER) is a critical task that aims to integrate cross-modal cues to identify human emotions. Prior MER work [17, 131, 56] can generally be grouped into two directions: improving modality fusion and improving feature representation. For fusion, early MER pipelines often adopted simple early-fusion baselines, where modality features are concatenated into a single representation [71]. Later work models cross-modal interactions explicitly, for example through tensor fusion that captures higher-order relationships between modalities [138]. Transformer-based approaches further use cross-modal attention to model interactions across unaligned multimodal sequences [118]. For feature representation, Self-MM uses self-supervised multi-task learning to generate unimodal supervision signals and learn stronger modality-specific representations without requiring additional unimodal labels [136]. However, these methods mainly formulate emotion recognition as a prediction task, often producing a final label or score without explaining which multimodal cues support the prediction [71]. More recently, the emergence of LLMs has supported a shift toward generative formulations for emotion

understanding, where models can not only predict emotions but also describe affective cues and generate human-understandable explanations [73].

2.5 LLM-Assisted Data Annotation

Recent work has begun to treat LLMs as part of the data work workflow, where LLMs collaborate on preprocessing and inspection, but still require human verification and explicit records of transformations [52]. Researchers explore LLMs as data preprocessors for tasks such as error detection, imputation, and matching, and highlight practical limitations that motivate careful checking [143]. Visual analytics research similarly argues that LLMs can help people navigate, query, and summarize data during interactive inspection, while stressing grounding and risk management rather than blind automation [41]. In annotation, Human–LLM collaborative workflows show that LLMs can propose labels and explanations while humans validate uncertain cases, improving efficiency without removing human control [122].

This direction is particularly relevant to multimodal emotion data annotation, which has increasingly moved from human emotion labels in datasets such as IEMOCAP [15] and CMU-MOSEI [139] toward richer formats in datasets such as MER-Caption [72] and OV-MERD [74], including natural language emotion descriptions and open-vocabulary labels. This shift makes annotation more demanding because analysts need to inspect temporally distributed cues, modality-specific, and contextual evidence together. LLMs can therefore support emotion dataset construction by generating candidate labels and description, with quality managed through human checks or model validation. For example, Jing et al. [48] apply GPT-4o with structured prompting to annotate a multimodal emotion dataset, and validate labels through human review and downstream training, showing both utility and the need for verification. Niu et al. [94] show that LLM emotion judgments can systematically differ from human annotations, and argue for integrating LLMs mainly for triage and quality control rather than fully replacing humans.

2.6 Using LLMs for Emotion Recognition

With the recent progress of LLMs, researchers have begun to explore their potential for emotion recognition [144]. Emotion-LLaMA [19] integrates visual, audio, and textual encoders with LoRA-based fine-tuning to improve multimodal affect recognition, while WisdoM [121] enhances sentiment classification by combining contextual world knowledge with multimodal cues. Another line of work applies prompt engineering, where zero-shot or few-shot prompts are designed to elicit emotional reasoning without

additional training. Studies demonstrated that task-specific and psychologically informed prompts can substantially improve emotion detection performance [2, 145]. Building on this, chain-of-thought reasoning has been introduced to enable step-by-step inference of affective cues. THOR [34] and ECR-Chain [40] show that structured reasoning helps LLMs identify implicit sentiments and emotion-cause relations. More recently, multi-agent prompting frameworks have been proposed to enhance robustness and interpretability, allowing multiple LLMs to act as emotional experts or evaluators that collaboratively refine predictions [126]. Together, these methods demonstrate that LLMs can serve as unified, explainable, and context-sensitive models for multimodal emotion reasoning.

2.7 Mixture-of-Experts

Mixture-of-Experts (MoE) architectures instantiate multiple parallel expert sub-networks for a given component and use a routing function to select or weight experts for each input [18, 43, 50]. This design is attractive for multimodal emotion recognition because different modalities can have different noise levels and time alignment quality, so the most reliable cues can vary across participants, scenes, and moments. Recent MER work uses MoE to perform sample-adaptive modality weighting and reduce negative transfer between modalities. For example, EMOE models modality-specific emotion experts with dynamic routing, and introduces modality-specific enhancement and distillation to preserve strong unimodal cues during fusion [32]. For continuous emotion prediction in more realistic settings, hierarchical MoE frameworks explicitly address incomplete and asynchronous inputs via multi-level routing and cross-modal alignment to reduce temporal mismatch [146]. Other designs, such as cross-attention gated MoE, use cross-modal attention to control expert contributions and encourage expert specialization, improving robustness when modalities provide conflicting evidence [91]. MoE has also been combined with structured fusion, for example multimodal graph attention with MoE-style expert weighting, to model relational context in interaction data rather than only feature concatenation [140]. Finally, some MoE-based frameworks incorporate human preference alignment to bias model outputs toward judgments that better match human ratings, which is relevant when objective physiological patterns and subjective affect reports do not fully agree [129].

CHAPTER 3

Understanding the Effects of Interaction on Emotional Experiences in VR

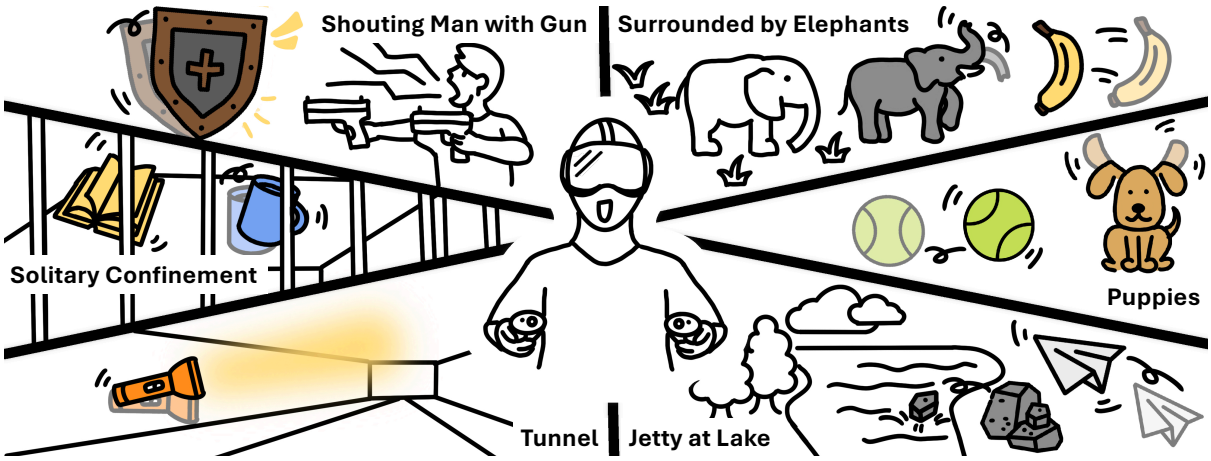


FIGURE 3.1. Illustration of the participant under six virtual reality scenes that could elicit emotions with interactive objects highlighted in color.

3.1 Scene Modeling

To understand how interaction affects emotion elicitation in VR, we extend the validated dataset of Jiang et al. [46] by incorporating interactive elements into each scene (details below). The original scene distribution was retained to preserve comparability with Jiang et al. [46] and to keep the affective variation among scenes, including different valence-arousal levels within the same quadrant. Each scene was designed to target a specific quadrant of Russell's circumplex model of affect [104] as shown in Figure 3.2: *Jetty at Lake* and *Puppies* fall into the Low Arousal High Valence (LAHV) quadrant, *Tunnel* and *Solitary Confinement* into the Low Arousal Low Valence (LALV) quadrant, and *Shouting Man with Gun* into the High Arousal Low Valence (HALV) quadrant. Furthermore, the original dataset did not include a scene targeting the High Arousal High Valence (HAHV) quadrant, which the authors acknowledged as a limitation due to ethical considerations and risks of cybersickness [46]. To address

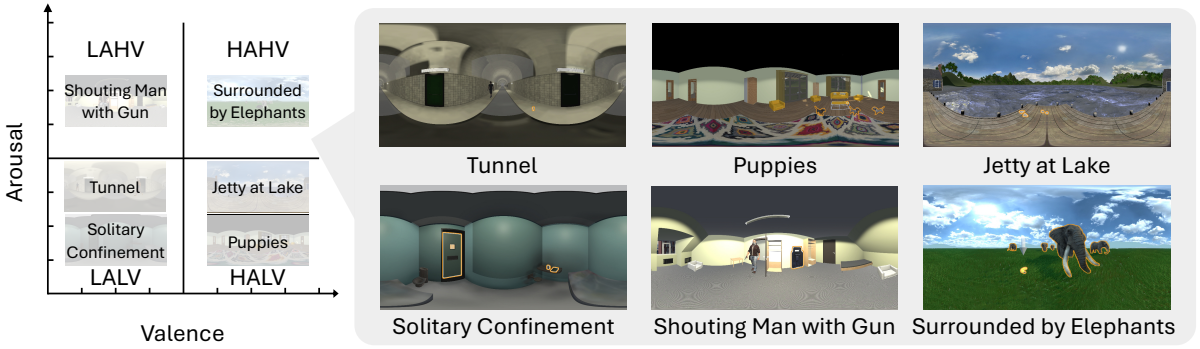


FIGURE 3.2. Overview of the six VR scenes. Left: placement of the scenes within Russell's valence–arousal circumplex model, covering all four quadrants (LAHV, HAHV, LALV, HALV). Right: panoramic views of the scenes with interactive objects highlighted in yellow.

these concerns, we ensured that the new HAHV scene was ethically appropriate, designed to elicit strong but positive emotions such as excitement and awe while avoiding distressing or fear content [77].

To address this gap, we followed Jiang et al. [46]'s approach and selected a HAHV video clip — *Surrounded by Elephants*[1] from Li et al. [65] consisting of 73 video clips. Following the setup of the original video, the user spawns in an open grassland under a bright sky, seeing a herd of seven elephants slowly approaching together. The scene offers a wide, unobstructed view, only limited by hills in the far distance. This video was selected to minimize the risk of cybersickness, as prior studies have shown that VR content with strong and dynamic visual motion (*e.g.*, roller coaster simulations) often provokes cybersickness [16], whereas gently paced, naturalistic scenes are expected to pose a lower risk of cybersickness. This consideration is a critical factor in mitigating potential discomfort and ensuring accessibility for a wide range of participants [46]. We then recreated this scene as an immersive VR environment by converting the 360° video into a 3D model using Blender and Unity.

Interactive Scene Design. To examine how interaction influences emotional responses, we designed the interactive elements to be intuitive and emotionally meaningful. Here, we define interaction as direct manipulation with objects within the scene (*e.g.*, picking up, throwing, touching, or using them). Each scene was implemented in two versions: an interactive version with user-triggered actions and a non-interactive version with identical audiovisual content, but no object-level interactions. This setup allows controlled comparisons of how interactive elements contribute to emotion elicitation.

[1] Link to the Surrounded by Elephants original video `https://www.youtube.com/watch?v=mlOiXMvMaZo`

In the interactive scenes, players can teleport and interact with game objects using controllers. All interactive elements in the scenes are highlighted with a yellow outline when the player's distance to the element is less than 2.5 meters, providing clear visibility and intuitive feedback about potential interactions in a natural reach or teleportation range [61, 45]. Each scene also includes corresponding sound effects as detailed in Table 3.1. As illustrated in Figure 3.2, the six modeled scenes highlight the interactive objects in yellow. The details of the interaction in six scenes are shown in Figure A.2 in Appendix A.3 and described below.

Tunnel. The player finds themselves in a long tunnel lit by dim lights, with pedestrians occasionally passing by. A flashlight lies on the tunnel floor. The player can pick it up. The action triggers light vibrotactile feedback in the controller. Once grabbed, the flashlight turns on and can be used to illuminate various areas of the tunnel.

Puppies. The player spawns in a spacious and furnished room with three puppies moving around. A tennis ball is placed on a table. The player can pet the puppies to trigger soft vibrotactile feedback, causing them to turn toward the player and sit down. The player can also pick up and throw the tennis ball; the puppies will chase the ball and bring it back. If the player does not interact with the puppies for over 15 seconds, the puppies will sit still on the ground.

Jetty at Lake. The player spawns on a jetty in front of a stone house by a calm lake, with hills covered by trees and grass. At the end of the jetty, two stones and two paper planes are placed. The player can grab the stones, triggering short vibrotactile pulse feedback, and throw them into the lake, which produces splash sounds and visible ripple effects on the water surface. Paper planes can also be picked up, triggering light vibrotactile feedback, and thrown, with a visible trajectory rendered at the tail during flight.

Solitary Confinement. The player spawns in a confined, gloomy cell with a flashing light, a toilet seat, and a single bed. A book and a metal cup are placed on the table. The player can knock on the iron door, triggering strong vibrotactile feedback and loud knocking sounds. The cup and book can be picked up and thrown at the door.

Shouting Man with Gun. The player spawns inside a furnished attic, where, after a short delay, a man breaks in shouting and aiming a pistol at them. A metal riot shield is placed against the wall, featuring

a small bulletproof glass window. The player can grab the shield for protection. The shield can be used to block the view, and the player can observe the man through the window.

Surrounded by Elephants. The player is placed in a green grassland with distant hills and a cloudy sky, where a herd of elephants slowly approaches. A banana floats above the grass. The player can grab the banana, trigger light vibrotactile feedback, and throw it toward the elephants. This triggers the elephant to pick it up with its trunk and eat it. When the player touches the elephant, it triggers deep vibrotactile feedback in the controller; the elephant takes two steps back, raises its trunk, and produces a vocal sound.

We developed our scenes using Unity 6000.0.46f1 Long-Term Support (LTS) version to ensure compatibility with the Meta Quest runtime environment and the SDKs employed in this study. All 3D models were either imported from open-source repositories licensed under CC BY-SA, obtained from the Unity Asset Store under its standard license, or created using Blender.

3.2 Study 1: Validating the effectiveness of the *Surrounded by Elephants* Scene

First, we carried out a user study following a within-subjects design to evaluate if the VR scene *Surrounded by Elephants* elicited HAHV emotions as effectively as its original 360° video format (RQ1). The study was approved by our university's research ethics committee.

3.2.1 Apparatus

The study took place in a university lab with an unobstructed room-scale tracking area of more than 3×3 metres. We used a Meta Quest Pro VR headset for the study. We rendered the scene in Unity and ran it on a desktop PC (Intel Core Ultra 9 285K, 32GB DDR5 RAM, Nvidia RTX 4080 SUPER) connected via a Quest Link cable.

TABLE 3.1. Details of the modeled interactive scenes for emotion elicitation. *V – Valence, A – Arousal*.

Scene Name	**Visual Description**	**Sound Effect**	**Interactive Elements**	**Targeted Emotion**	**Ref.**
Tunnel	A long tunnel lit by dim lights, with pedestrians occasionally passing by.	Footsteps	Picking up a flashlight.	V: Mid-to-low A: Mid-to-low	[108]
Puppies	A spacious and furnished room with several puppies around.	(Quiet)	Petting puppies and throwing ball.	V: High A: Low	[108, 65]
Jetty at Lake	A jetty in front of a stone house by a lake, with hills covered by trees and grass.	Water flow	Throwing stones and paper airplanes.	V: High A: Low	[108]
Solitary Confinement	A gloomy cell with a flashing light, a toilet set, and a single bed.	Water dropping	Knocking door; grabbing a book and a cup.	V: Low A: Low	[65]
Shouting Man with Gun	A furnished attic. A furnished attic where a man breaks in while shouting and aims a pistol at the player after a certain time.	Man shouting	Using shield to block and observe.	V: Mid-to-low A: High	[108]
Surrounded by Elephants	A green grassland with distant hills and a cloudy sky, where elephants walk toward.	Wind and elephant trumpeting	Feeding and touching elephants	V: High A: High	[65]

3.2.2 Procedure

Upon arrival in our lab, we assigned each participant a unique anonymous identifier for data management. Moreover, we screened participants for potential health risks related to VR use (*e.g.*, epilepsy, mobility or visual impairments, adverse emotional responses). Eligible participants received an overview of the study, provided informed consent, and completed a questionnaire on demographics and prior VR experience. Participants then completed a tutorial scene to familiarize themselves with the VR system, during which a baseline SAM questionnaire [14] was administered. Participants subsequently experienced one 360° video and one VR scene in a balanced random order, each lasting at least 30 seconds with the option to continue exploring before completing the SAM scale (see Figure A.1). Between

scenes, participants transitioned back to the lab environment via the headset's mixed reality feature for a 60-second break. The study lasted approximately 10 minutes per participant.

3.2.3 Participants

We recruited 24 volunteers (12 female, 12 male) aged between 19 and 38 years (M = 25.63, SD = 3.70) to participate in our study. All participants provided informed consent prior to the study. They did not receive monetary compensation, but were offered snacks. Our participants reported different levels of VR experience: 8 used VR daily, 7 weekly, 5 monthly, and 4 never.

3.2.4 Results

We first evaluated how effectively the *Surrounded by Elephants* VR scene elicited emotion. We collected a total of 72 SAM measurements (24 participants $\times$ 2 CONDITIONS $\times$ *Surrounded by Elephants*). The two CONDITIONS were the original 360° video [65] and the respective modeled VR scene.

After checking for normality, we applied ART ANOVAs [127] to analyze the non-normally distributed data (Shapiro-Wilk test, $p < 0.05$). For Valence, the main effect of CONDITION was significant, $F(1, 23) = 15.597$, $p < 0.001$, $\eta^2 = 0.440$, indicating that the *Surrounded by Elephants* VR scene elicited significantly higher ratings than the 360° Video. For Arousal, the main effect of CONDITION was not significant, $F(1, 23) = 0.055$, $p = 0.817$, $\eta^2 = 0.004$, indicating no difference between the VR Scene and the 360° Video. For Dominance, the main effect was significant, $F(1, 23) = 9.015$, $p < 0.01$, $\eta^2 = 0.368$, with higher ratings for the VR scene compared to the 360° video. Overall, the VR scene elicited higher valence and similar arousal compared to its 360° video counterpart, supporting its use as a high-valence, high-arousal VR scene. The VR scene also elicited higher dominance, suggesting that participants felt a stronger sense of control in the modeled VR environment, consistent with previous findings. Our results show that our *Surrounded By Elephants* VR Scene can elicit high-valence, high-arousal emotion. Table A.2 in Appendix A.1 presents the results, showing that the VR scene elicited emotional responses comparable to the 360° video, indicating its effectiveness for high-arousal, high-valence emotion elicitation.

FIGURE 3.3. An example of the experimental setup for participant experiencing VR scenes.

3.3 Study 2: Understanding Effects of Interaction on Emotion Elicitation

Then we conducted a mixed-design user study to investigate the role of affective interaction when eliciting emotions in VR (RQ2 and RQ3). Affective interaction design (Interactive vs. Non-Interactive) served as a between-subject factor, while the six VR scenes were used as a within-subject factor. We applied a balanced Latin square to counterbalance the scene order across participants. Physiological and self-report measures were collected throughout to examine how interactive elements influence emotional responses and how these responses are reflected in physiological signals. This user study received ethics approval from our university.

3.3.1 Apparatus

We used a Meta Quest Pro VR headset and two Meta Quest Touch Pro controllers for all VR scenes. The study was conducted in a university lab space with an unobstructed room-scale tracking area of over 3×3 meters to support full-body interaction. All scenes were rendered in Unity and ran on a desktop PC (Intel Core Ultra 9 285K, 32GB DDR5 RAM, Nvidia RTX 4080 SUPER) via a Quest Link cable connection. The experimental setup is shown in Figure 3.3.

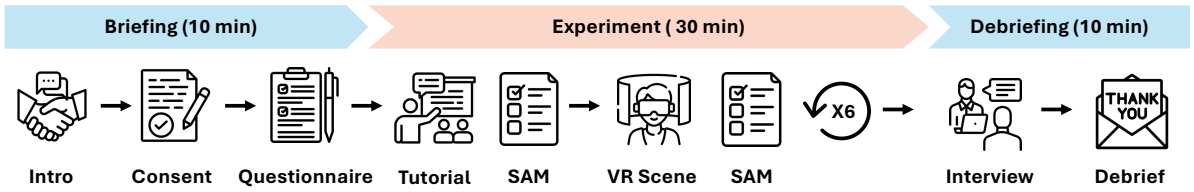


FIGURE 3.4. Study Protocol: Some study elements were conducted outside the VR environment (blue), while the main study components occurred within the VR environment to avoid breaking participants' immersion (orange).

3.3.2 Measurements

During the study, we used the SAM questionnaire to assess participants' emotional states. We also collected their physiological data using the Empatica EmbracePlus wristband during the study. We recorded electrodermal activity (EDA) and blood volume pulse (BVP) signals via the Empatica Care Lab mobile application. EDA was sampled at 4 Hz, and BVP at 64 Hz. In addition, we conducted semi-structured interviews in which participants identified the scenes they found most and least emotionally impactful and provided explanations for their choices.

3.3.3 Participants

We recruited 84 participants (42 female, 42 male) aged between 19 to 35 years (M = 24.05, SD = 3.03). All participants provided informed consent prior to the study and were compensated with a gift card ($\approx 15USD$) for their participation. Each participant was assigned a unique anonymous identifier for data management. Participants came with various VR usage experiences, with 14 reporting daily use, 21 weekly, 25 monthly, and 24 never. Table A.1 in Appendix A.1 presents the demographic details for the user study.

3.3.4 Procedure

The study protocol is illustrated in Figure 3.4. Upon arrival in our lab, participants were first screened for potential health risks related to VR use, including epilepsy, mobility impairments, severe visual impairments, and a history of adverse emotional responses such as anxiety disorders or PTSD through a questionnaire. We then provided eligible participants with an overview of the study's purpose. After confirming their understanding, they provided written consent and completed a questionnaire that collected demographic information, including gender, age and prior VR experience. We then fitted the

participants with an EmbracePlus wristband to record physiological data, which they wore for the entire duration of the study, approximately 50-60 minutes per participant. After that, we asked participants to complete a tutorial, designed to introduce them to the VR system and ensure their familiarity with the headset and control methods, including teleportation for navigation and interaction with the designated virtual objects.

The tutorial scene was implemented in two versions to familiarize participants with the VR system. Both versions were set in an open, unobstructed virtual space with a text panel in front of the participant explaining the study procedure and controls. In the non-interactive version, participants practiced teleportation by moving to an exit point and then completed the SAM questionnaire [14]. In the interactive version, two blue cubes were placed in front of the participant, and participants practiced picking them up and throwing them before teleporting to the exit point to complete the SAM questionnaire. Within this scene, we also collected a baseline measurement of their emotional state using the SAM questionnaire.

Once comfortable with the setup, participants proceeded to experience all six VR scenes in a counterbalanced order using a Latin square design. Each scene lasted for at least 30 seconds, after which an exit point appeared. Participants could, however, choose to remain in the scene for further exploration or leave the scene by teleporting to the exit point. Upon leaving, they were teleported to the initial position of the scene to complete the SAM questionnaire. Afterwards, they exited the scene and returned to the mixed reality environment of the lab for a 60-second break to reduce any potential emotional carryover.

After completing all scenes, participants exited the VR application and took part in a semi-structured interview. This final phase aimed to gather qualitative insights into their emotional responses and overall experiences with each scene.

3.4 Evaluation Results

To capture how object-level interaction shapes emotional experience in VR, we detail the findings of our study in this section and provide a comprehensive evaluation of the interaction of the scenes. Precisely, we first compared SAM results between the interactive and non-interactive versions to examine differences in emotional elicitation. To uncover the behavioral and contextual mechanisms behind these ratings and gain a more comprehensive understanding of the emotion elicitation process, we further analyzed participants' interactions with the virtual environment, including *time* engaged with each scene, *virtual position* and *orientation*, and emotion-related *descriptions* extracted through topic modeling. In

addition, we analyzed physiological data, including BVP and EDA, to complement the assessment of emotional responses.

3.4.1 Emotion Elicitation Measurements

We analyzed 588 SAM measurements (42 participants per group $\times$ 2 AFFECTIVE INTERACTION DESIGN $\times$ SCENE (6 scenes, and the tutorial as baseline)). The AFFECTIVE INTERACTION DESIGN refers to the interactive and non-interactive versions of the VR scenes. As the data deviated from normality (Shapiro–Wilk test, $p < .05$) and the study used a mixed design with repeated measurements across scenes, we used linear mixed-effects models to analyze SAM ratings, focusing on between-group effects. The model specified AFFECTIVE INTERACTION DESIGN and SCENE as fixed effects with their interaction, included Gender, Age, and XR experience as control variables, and added a random intercept for participants to account for within-subject dependency. The fixed-effects estimates are summarized in Table 3.2.

The emotional responses in the valence-arousal space are illustrated in Figure 3.5, Figure 3.6, and Figure 3.7. The results indicate differences in emotional responses based on the presence of interactive elements. For example, in the *Puppies* scene, the interactive version elicited higher valence, arousal, and dominance ratings than the non-interactive version. In the *Shouting Man with Gun* scene, interaction led to higher valence and dominance, while arousal remained similarly high across both versions.

3.4.1.1 Effects of interaction

As illustrated in Table 3.2, across all scenes, the main effect of the interaction was not significant: Valence (Estimate = -0.01, SE = 0.36, 95% CI = [-0.70, 0.71]), Arousal (Estimate = 0.06, SE = 0.47, 95% CI = [-0.87, 0.98]), and Dominance (Estimate = -0.29, SE = 0.44, 95% CI = [-1.15, 0.57]). By contrast, the effects of the scene were robust, eliciting the targeted emotions, *e.g.*, reduced Valence and increased Arousal in *Shouting Man with Gun*, or decreased Valence and Dominance in *Solitary Confinement*.

The interaction between AFFECTIVE INTERACTION DESIGN and SCENE revealed how affective interaction influenced experiences in specific contexts. Valence increased in *Puppies* (Estimate = 0.88, SE = 0.42, 95% CI = [0.05, 1.71]), with participants rating higher values when they could pet and play the ball with the puppies. Arousal decreased in *Tunnel* (Estimate = -1.14, SE = 0.55, 95% CI = [-2.23,

TABLE 3.2. Fixed-effects estimates from linear mixed-effects models for Valence, Arousal, and Dominance. Each outcome was modeled as *Outcome* $\sim$ Affective Interaction Design * Scene + Gender + Age + XR Experience + (1|Participant), where *Affective Interaction Design* refers to interactive vs. non-interactive, and *Scene* includes one tutorial (baseline) and six VR scenes. Values are reported as Estimate (SE) with 95% CI.

Parameter	**Valence**			**Arousal**			**Dominance**		
	Estimate			Estimate			Estimate		
	M	SE	95% CI	M	SE	95% CI	M	SE	95% CI
Intercept	**6.37*****	0.92	[4.54, 8.20]	**5.80*****	1.21	[3.39, 8.20]	**7.58*****	1.13	[5.33, 9.83]
Interactive	-0.01	0.36	[-0.71, 0.70]	0.06	0.47	[-0.87, 0.98]	-0.29	0.44	[-1.15, 0.57]
Gender (Male)	0.26	0.22	[-0.18, 0.70]	**-0.85****	0.29	[-1.43, -0.27]	0.46	0.27	[-0.09, 1.00]
Age	0.02	0.04	[-0.05, 0.09]	-0.04	0.05	[-0.13, 0.06]	-0.02	0.05	[-0.11, 0.07]
VR Experience	-0.07	0.11	[-0.28, 0.14]	0.26	0.14	[-0.02, 0.53]	-0.24	0.13	[-0.50, 0.01]
Tunnel	**-1.76*****	0.3	[-2.35, -1.17]	0.26	0.39	[-0.51, 1.03]	**-2.21*****	0.36	[-2.93, -1.50]
Puppies	0.19	0.3	[-0.40, 0.78]	-0.43	0.39	[-1.20, 0.34]	-0.64	0.36	[-1.35, 0.07]
Jetty at Lake	-0.07	0.3	[-0.66, 0.52]	-0.43	0.39	[-1.20, 0.34]	-0.31	0.36	[-1.02, 0.40]
Solitary Confinement	**-2.76*****	0.3	[-3.35, -2.17]	-0.50	0.39	[-1.27, 0.27]	**-2.81*****	0.36	[-3.52, -2.10]
Shouting Man with Gun	**-2.60*****	0.3	[-3.18, -2.01]	**2.05*****	0.39	[1.28, 2.81]	**-2.86*****	0.36	[-3.57, -2.15]
Surrounded by Elephants	-0.02	0.3	[-0.61, 0.56]	**1.00***	0.39	[0.23, 1.77]	**-1.17****	0.36	[-1.88, -0.46]
Interactive $\times$ Tunnel	0.02	0.42	[-0.81, 0.86]	**-1.14***	0.55	[-2.23, -0.06]	**1.48****	0.51	[0.47, 2.48]
Interactive $\times$ Puppies	**0.88***	0.42	[0.05, 1.71]	0.29	0.55	[-0.80, 1.37]	**1.74*****	0.51	[0.73, 2.74]
Interactive $\times$ Jetty at Lake	0.67	0.42	[-0.17, 1.50]	-0.48	0.55	[-1.56, 0.61]	**1.05***	0.51	[0.04, 2.05]
Interactive $\times$ Solitary Confinement	0.40	0.42	[-0.43, 1.24]	0.62	0.55	[-0.47, 1.70]	**1.14***	0.51	[0.14, 2.15]
Interactive $\times$ Shouting Man with Gun	0.64	0.42	[-0.19, 1.47]	-0.19	0.55	[-1.28, 0.89]	0.83	0.51	[-0.17, 1.84]
Interactive $\times$ Surrounded by Elephants	0.21	0.42	[-0.62, 1.05]	-0.26	0.55	[-1.35, 0.82]	**1.12***	0.51	[0.11, 2.12]

Notes. Stars denote significance: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

-0.06]) when participants could use the flashlight. Dominance increased in *Puppies* (Estimate = 1.74, SE = 0.51, 95% CI = [0.73, 2.74]), *Tunnel* (Estimate = 1.48, SE = 0.51, 95% CI = [0.47, 2.48]), *Solitary Confinement* (Estimate = 1.14, SE = 0.51, 95% CI = [0.14, 2.15]), *Jetty at Lake* (Estimate = 1.05, SE = 0.51, 95% CI = [0.04, 2.05]), and *Surrounded by Elephants* (Estimate = 1.12, SE = 0.51, 95% CI = [0.11, 2.12]). Figure 3.8 illustrates these differences, showing that interaction did not uniformly shift responses, but produced scene-dependent effects, most consistently in measures of Dominance, i.e., the sense of control a user feels over a situation during an emotional experience. Covariates further revealed that male participants reported lower Arousal (Estimate = -0.85, SE = 0.29, 95% CI = [-1.43, -0.27]) than female participants.

Figure 3.5 shows scene-specific differences. Interaction generally yielded higher valence and dominance in *Puppies*, *Surrounded by Elephants*, *Jetty at Lake* scenes. In these scenes, participants engaged more, petting the puppies and playing fetch with them, feeding the elephants, and throwing stones or paper planes. Differences were small or absent in *Shouting Man with Gun* and *Solitary Confinement*. For arousal, effects of interaction were mixed, with lower values in *Tunnel*, but with little or no change in other scenes.

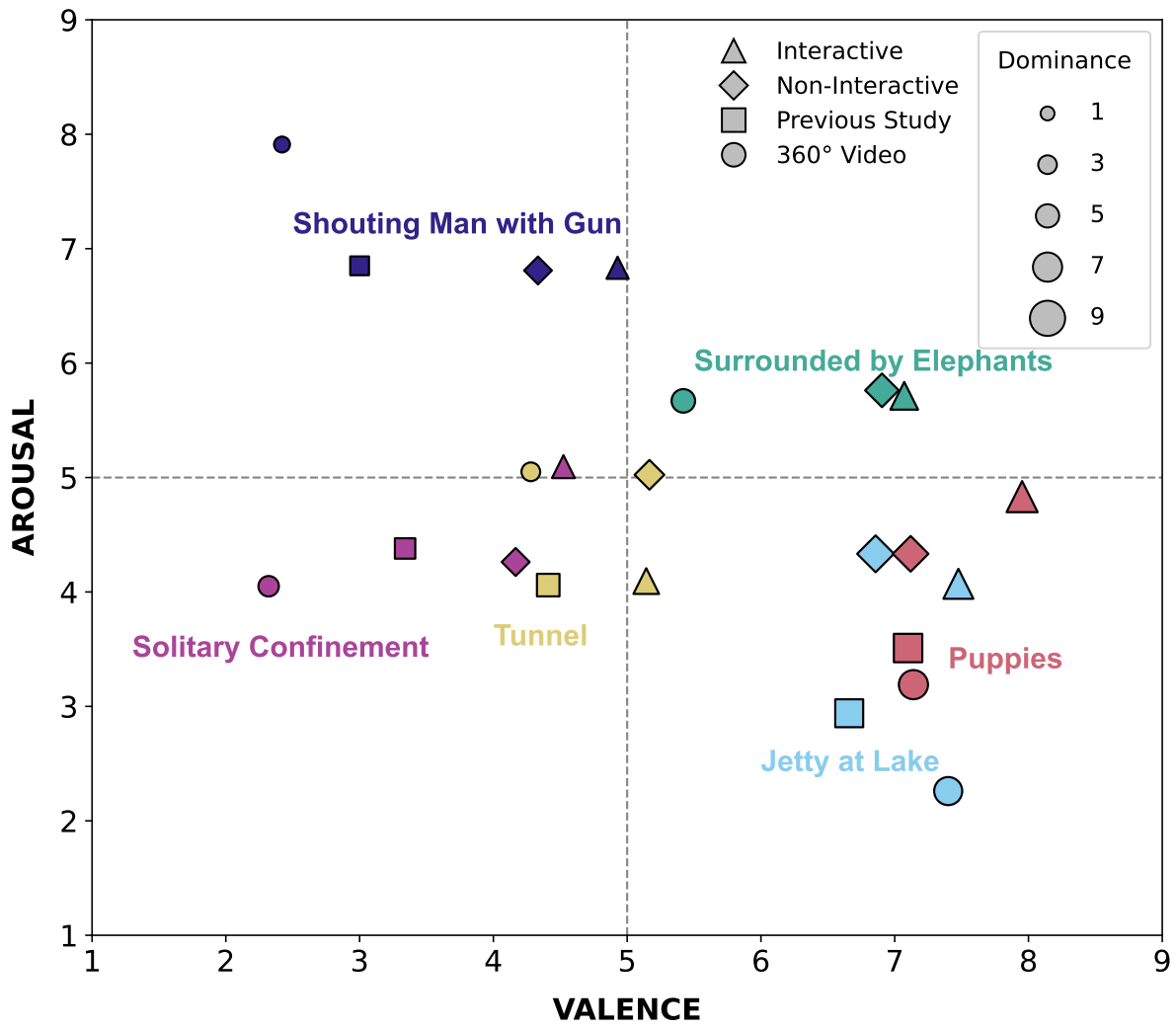


FIGURE 3.5. Illustrations of all emotional responses in the valence-arousal space. The x-axis represents valence, the y-axis represents arousal, and marker size represents dominance. Interactive and Non-Interactive results are from this study, Previous Study results from Jiang et al. [46], and 360° results from established datasets [65, 108]. The *Surrounded by Elephants* scene is not included in the Previous Study.

Taken together, these patterns indicate that affective interaction design consistently elevated Dominance across different contexts, allowing participants to feel more in control of the emotional experience rather than passively receiving it. In contrast, its impact on Valence and Arousal varied across scenes, suggesting that interaction does not simply amplify emotions uniformly, but rather shapes the user's experience in a context-dependent manner.

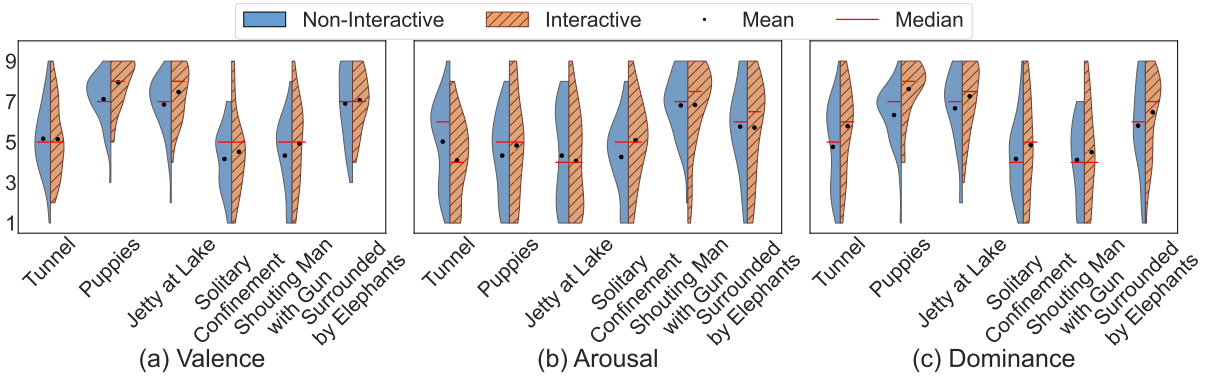
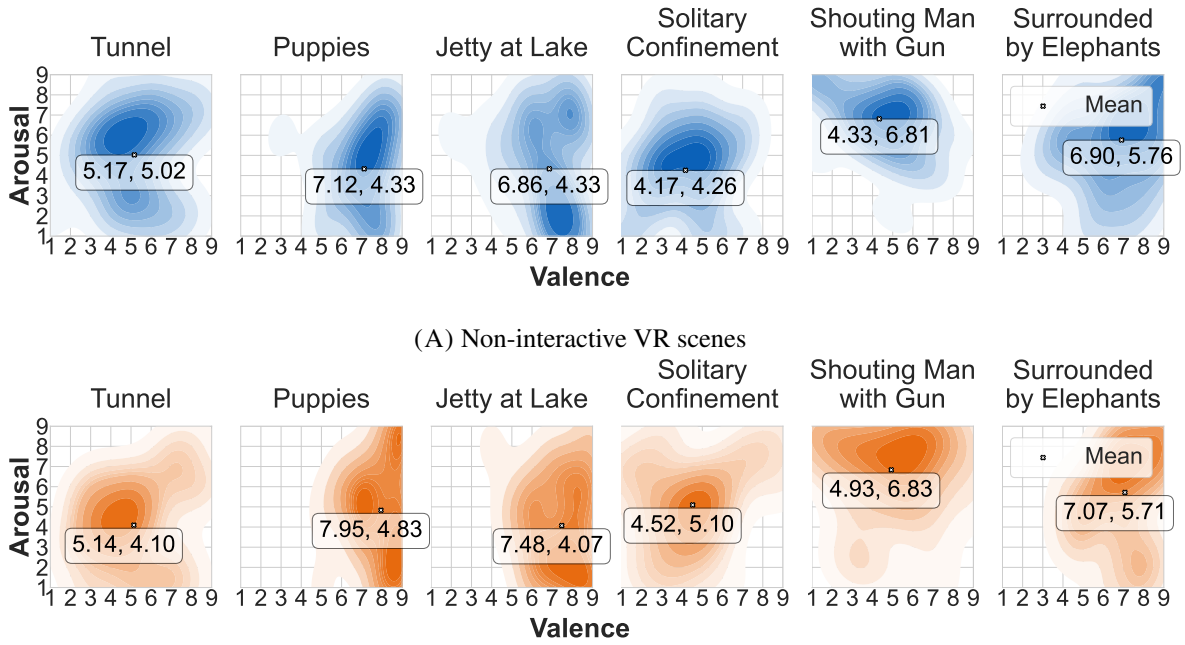


FIGURE 3.6. SAM questionnaire results after experiencing each scene for valence, arousal, and dominance.



(A) Non-interactive VR scenes

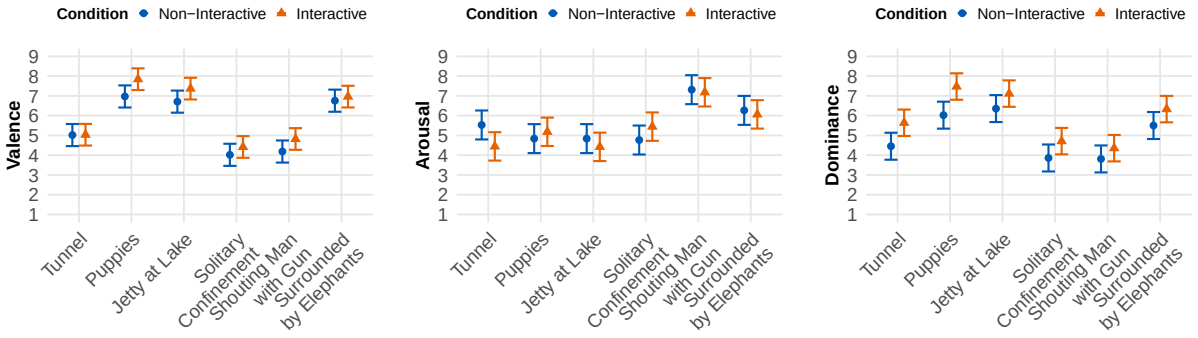
(B) Interactive VR scenes

Kernel density estimate plots of valence and arousal values for interactive VR scenes.

FIGURE 3.7. Kernel density estimate plots of valence and arousal values.

3.4.2 Engagement Time in VR Scenes

Based on the SAM results, we assess user engagement time to understand how interaction influences the emotion-elicitation process. We compared the time participants spent in each scene under the interactive and non-interactive conditions (see Figure 3.9 for details). Engagement time (in seconds, error: ± 0.1 s) was measured from the moment participants entered the scene until participants exited it. Following Jiang et al. [46]'s approach, a minimum engagement time of 30 seconds was set to provide consistent



(A) Valence (B) Arousal (C) Dominance

FIGURE 3.8. Model-predicted interaction effects from the mixed-effects models for (A) Valence, (B) Arousal, and (C) Dominance across interactive and non-interactive conditions. Each point represents the estimated marginal mean for a given scene, with error bars showing the 95% confidence interval.

exposure across scenes before participants could exit. The Shapiro–Wilk test indicated that engagement time distributions deviated from normality ($p < 0.05$). We thus used a linear mixed-effects model with participant as a random intercept to compare scenes across both conditions, revealing significant differences in mean engagement time.

The linear mixed-effects model revealed a significant main effect of interaction ($p < 0.001$), with participants spending longer in interactive than in non-interactive scenes. This effect was most pronounced in the *Puppies* scene (interactive: $M = 133.58 \pm 89.92$ s, $Med = 113.87$ s; non-interactive: $M = 49.73 \pm 21.46$ s, $Med = 42.92$ s), where participants stayed more than twice as long on average when interaction was enabled. Substantial differences were also found in *Jetty at Lake* ($M = 73.39$ s vs. 41.10 s), *Surrounded by Elephants* ($M = 89.77$ s vs. 52.66 s), and *Solitary Confinement* ($M = 75.05$ s vs. 39.93 s), indicating that interaction extended engagement compared to the non-interactive scenes.

Overall, these results show that interaction significantly prolonged user exposure to the stimuli, suggesting that the opportunity to act encourages participants to sustain their engagement rather than exiting early. This indicates that affective interaction designs provide participants with greater opportunities to fully experience the intended emotional context within the virtual environment.

3.4.3 Engagement with Virtual Environments

To investigate how interaction shapes participants' spatial engagement and how these behavioral patterns correlate with emotional responses, we analyzed participants' engagement with the virtual environments

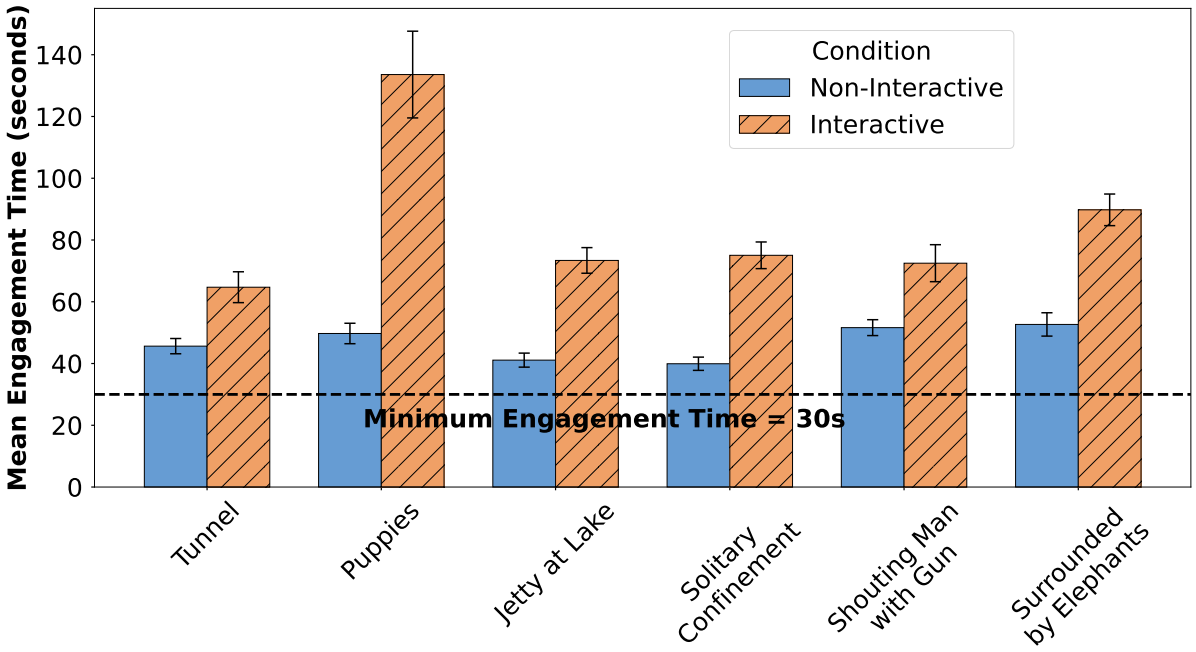


FIGURE 3.9. Mean engagement time across scenes for interactive and non-interactive conditions.

by comparing their virtual positions and orientations between the interactive and non-interactive conditions. For each scene, we identified the areas within the scenes where participants spent the most time. The analysis focused on the XZ plane, representing the top-down view of positions in the virtual space (in Unity, the Y dimension corresponds to height). Figure 3.10 presents the sampled position data for each condition across scenes. By examining these spatial engagement patterns in combination with topic modeling results, we identified the following differences between the two conditions:

Tunnel. In the non-interactive condition, participants largely remained near the entrance of the corridor. With interaction enabled, participants advanced deeper into the tunnel after picking up the flashlight and explored the side doors, resulting in a more elongated spatial distribution. These behaviors were associated with higher dominance and lower arousal ratings in the interactive condition.

Puppies. Participants in the non-interactive condition mainly clustered near the entrance to observe the puppies. Under the interactive condition, they moved more widely across the room, engaging with the environment by petting the puppies or playing fetch using a tennis ball. This wider exploration corresponds to the higher valence, arousal, and dominance ratings.

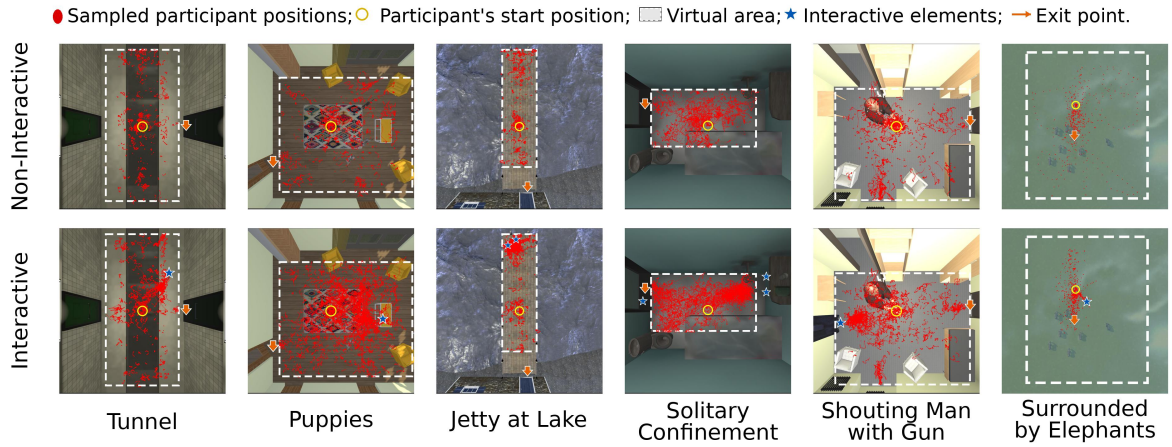


FIGURE 3.10. Spatial engagement patterns across the six VR scenes under non-interactive and interactive conditions. The positions are sampled every 0.1 seconds (sampling frequency = 10 Hz).

Jetty at Lake. Movements in the interactive condition extended toward the end of the jetty, where stones and paper planes were located, creating clusters there. Compared to the limited exploration in the non-interactive condition, these behaviors were associated with higher valence and dominance as well as reduced arousal.

Solitary Confinement. Across both conditions, due to the nature of the scene, movement was highly restricted, with dense clusters near the spawning point. In the interactive condition, participants moved slightly toward the table and door to interact with the book, cup, or knocking on the door, resulting in a modest expansion of spatial distribution. These behaviors were accompanied with higher dominance and arousal, though valence remained low.

Shouting Man with Gun. In the non-interactive condition, participants often moved toward the window at the back of the room. By contrast, the interactive condition involved more frequent approaches to the riot shield along the left wall and to the gunman himself. This shows more active coping behaviors and corresponds to higher dominance, whereas both conditions showed low valence and high arousal.

Surrounded by Elephants. In the non-interactive condition, participants dispersed toward the periphery of the elephant herd, producing a wider spatial spread. With interaction, movements concentrated within the herd's range, focusing on feeding bananas or touching elephants. These interactions yielded

higher valence, lower arousal, and greater dominance, reflecting a more positive and controlled emotional experience.

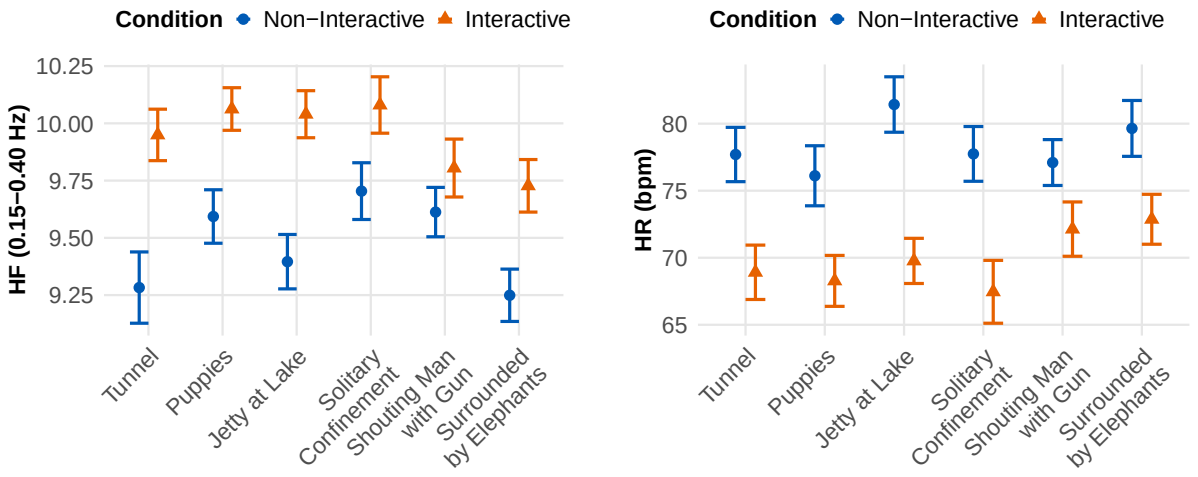
Overall, the spatial engagement analysis shows that interactive elements encouraged participants to move beyond the spawning point and interact with salient features of the environment. This increased exploration was most consistently associated with higher dominance, while changes in valence and arousal varied depending on the emotional character of each scene. Collectively, these spatial patterns demonstrate that interactive elements function as attentional anchors, effectively guiding users' physical movement and focus toward key emotional features of the scene.

3.4.4 Physiological Measures

In addition to the before-mentioned measures, we also analyzed physiological responses recorded with the EmbracePlus wristband and processed the signals using `NeuroKit2` [83]. From the raw BVP data, we derived Heart Rate Variability (HRV) and Heart Rate (HR). From the raw EDA data, we extracted the phasic component (Skin Conductance Responses, SCRs), SCR frequency, and mean SCR amplitude as indicators of arousal [13, 4].

From BVP, we detected systolic peaks and derived inter-beat interval (IBI) as the basis for HRV and HR computation [88]. Following the general HRV analysis procedure [88, 106], we interpolated and resampled the IBI data at 4 Hz to align the signals in a uniform time interval, and removed long-term trends through detrending. We then obtained the HRV power spectrum from the detrended IBI data by applying a short-time Fourier transform (STFT) with a 64-sized sliding window. Within each window, we calculated the High-Frequency (HF: 0.15–0.40 Hz) component of HRV and applied a natural log transformation to normalize the distribution. We also computed the mean HR within each segment. Higher HF HRV is typically interpreted as stronger parasympathetic activation, which shows lower physiological arousal [3]. Similarly, lower HR also indicates reduced arousal [3].

We applied linear mixed-effects models to the measures extracted from BVP signals, which revealed significant effects of interaction. Estimates are reported in Table A.3 in Appendix A.1, and Figure 3.11A and Figure 3.11B. HF HRV was significantly higher in the interactive condition (Estimate = 0.69, SE = 0.17, 95% CI [0.36, 1.02]), while mean HR was significantly lower (Estimate = -9.25, SE = 2.82, 95% CI [-14.80, -3.70]). Scene contrasts further showed increased HF HRV in *Puppies* (Estimate = 0.31, SE = 0.12, 95% CI [0.06, 0.55]), *Solitary Confinement* (Estimate = 0.44, SE = 0.13, 95% CI [0.19, 0.69]), and



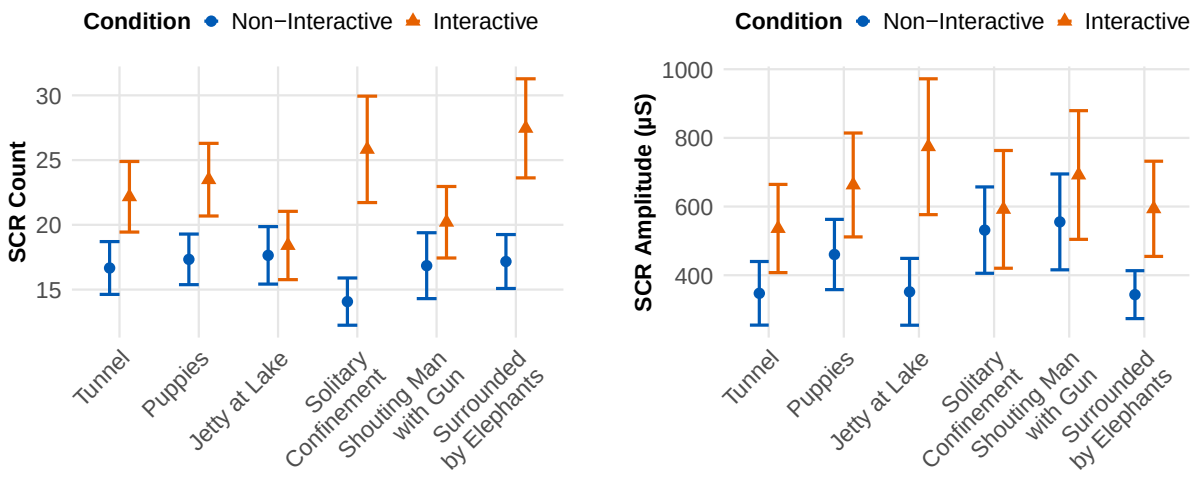
(A) HF (B) HR

FIGURE 3.11. Fixed-effects estimates for (A) HF and (B) HR.

Shouting Man with Gun (Estimate = 0.30, SE = 0.13, 95% CI [0.05, 0.55]). Notably, in *Shouting Man with Gun*, the increase in HF HRV associated with interaction was significantly attenuated (Estimate = -0.49, SE = 0.18, 95% CI [-0.84, -0.14]), suggesting that interaction may have limited regulatory benefit in strongly negative contexts. This pattern is in line with the SAM ratings reported for these scenes.

We analyzed the phasic component of skin conductance responses (SCR). Following established procedures for EDA analysis [13], the raw signals were filtered and deconvolved to separate phasic activity from the tonic baseline. For each segment, we quantified two features: SCR count, representing the total number of responses, and mean SCR amplitude, representing the average response magnitude. Higher SCR count and amplitude are typically interpreted as stronger sympathetic activation, which reflects greater physiological arousal [13].

The linear mixed-effects models results for SCR count and SCR amplitude are shown in Table A.4 in Appendix A.1 and Figure 3.12A and 3.12B. Although no statistically significant main effects of interaction or scene-level interactions were observed, consistent directional trends emerged that suggest increased sympathetic activation in the interactive condition. Specifically, for SCR count, marginal increases were observed across most scenes, with the largest differences in *Tunnel*, *Puppies*, *Solitary Confinement*, and *Surrounded by Elephants*. For SCR amplitude, a similar pattern was found, with stronger responses when interaction in the scenes was enabled, most prominently in *Jetty at Lake* and *Surrounded by Elephants*. These observations suggest that interaction elicited more frequent and stronger sympathetic responses.



(A) SCR Count (B) SCR Amplitude

FIGURE 3.12. Fixed-effects estimates for (A) SCR Count and (B) SCR Amplitude.

In summary, the physiological results show that interaction led to significantly higher HF HRV and lower HR compared to the non-interactive scenes, while simultaneously eliciting consistent trends toward more frequent and stronger SCRs. This suggests that interaction creates a composite physiological state, characterized by overall relaxation interspersed with moments of high reactivity.

3.4.5 Topic Modeling

We further examined the emotion-elicitation process through participants' qualitative descriptions collected in semi-structured interviews. Because emotions are inherently subjective, this analysis complements our quantitative results by capturing how participants themselves articulated their affective states. In particular, we focused on how these accounts differed between interactive and non-interactive scenes, providing richer insights into the role of interaction in shaping emotional experiences.

When analyzing the interview data, we applied topic modeling [46, 82]. Transcripts were first preprocessed by retaining nouns and adjectives related to emotional qualities, while stop words and low-information terms (e.g., feeling, scene, little) were removed. All texts were lemmatized to normalize word variations (e.g., plurals to singular). We then used Latent Dirichlet Allocation (LDA) [12] to extract three latent topics per scene, each represented by sets of co-occurring words that captured common

themes in participants' narratives [82]. For interpretation, we manually examined the five most representative words for each topic and visualized overall word distributions with word clouds to illustrate how interactive elements shaped emotional expressions across different scenes.

3.4.5.1 Topics: Tunnel

The Tunnel scene placed participants in a long corridor with dim yellowish lighting and occasional pedestrians. In the non-interactive condition, it elicited mid valence ($M = 5.17 \pm 1.71$, $Med = 5.00$), mid arousal (M = 5.02 ± 2.05, Med = 6.00), and neutral dominance (M = 4.76 ± 2.08, Med = 5.00). Through participants' descriptions, as illustrated in Figure 3.13, we observed mixed emotions described as "oppressive" (N = 3), "uncanny" (N = 2), "scary" (N = 2), "strange" (N = 2), and "frightening" (N = 2). LDA further highlighted themes of fear, oppression, and the unsettling presence of passersby. These patterns were often associated with the tunnel's confined structure, as one participant noted: *"There was an uncanny feeling, and the environment was dim, which made me feel a bit scared."* (P67). Others emphasized the unsettling effect of pedestrians, for example: *"The passersby made me feel a bit scared." (P55) and "It felt frightening, and the people seemed strange."* (P65).

In the *interactive version*, when participants were given a flashlight, the scene instead elicited mid valence (M = 5.14 ± 1.79, Med = 5.00), lower arousal (M = 4.10 ± 2.06, Med = 4.00), and higher dominance (M = 5.79 ± 2.19, Med = 6.00). Topic modeling revealed salient words such as "people" (N = 6), "flashlight" (N = 4), "scared" (N = 4), and "creepy" (N = 2). Topic modeling (LDA) indicated that while elements of eeriness remained, participants emphasized the flashlight as a source of agency and reassurance, reducing uncertainty about people's behavior. For instance, one participant stated: *"The confined space and low brightness of the colors made me concerned about people's behavior. I wanted to leave. Using the flashlight to see clearly showed that people were just walking normally, and their casual clothing reduced the sense of pressure."* (P9). Another participant explained: *"The environment was dark, and using the flashlight to light up the face made it less frightening."* (P17).

Overall, the Tunnel scene evoked unease through its narrow structure and the movement of pedestrians. In the non-interactive condition, participants emphasized feelings of fear, strangeness, and lack of control. In contrast, in the interactive condition, participants reported that the flashlight afforded agency, making the experience less threatening by reducing arousal and enhancing dominance.

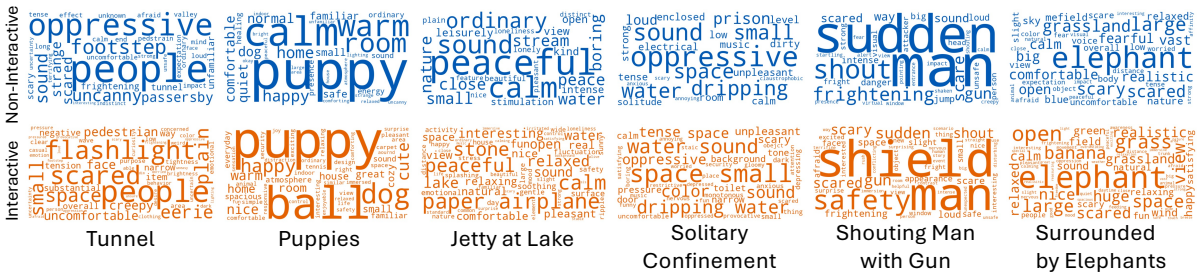


FIGURE 3.13. Word clouds of participants' descriptions of six VR scenes (top: Non-interactive, bottom: Interactive).

3.4.5.2 Topics: Puppies

The *Puppies* scene elicited emotions with high valence (M = 7.12 ± 1.25, Med = 7.00), low arousal (M = 4.33 ± 2.02, Med = 5.00), and high dominance (M = 6.33 ± 1.48, Med = 7.00) in the non-interactive condition. As illustrated in Figure 3.13, participants frequently described the experience as relaxing and delightful, often using words such as "puppy" (N = 8), "warm" (N = 4), "room" (N = 3), and "happy" (N = 3). LDA showed that these responses clustered around themes of familiarity, comfort, and calmness, with puppies and the indoor setting reinforcing feelings of safety and warmth: *"The puppies made me happy, and the indoor environment felt very familiar and everyday"* (P67).

When *interaction was enabled*, participants could pet and play with the dogs, which heightened their affective engagement. Under this condition, the scene was rated with even higher valence (M = 7.95 ± 1.17, Med = 8.00), slightly greater arousal (M = 4.83 ± 2.62, Med = 5.00), and higher dominance (M = 7.62 ± 1.43, Med = 8.00). Topic modeling highlighted terms such as "puppy" (N = 16), "ball" (N = 12), "dog" (N = 8), "happy" (N = 8), and "cute" (N = 6). LDA indicated themes of joy, playfulness, and comfort, emphasizing how active interaction with the puppies created both excitement and warmth. Participants highlighted the enjoyment of immersive engagement: *"I enjoyed that when I threw the ball, the dogs brought it back."* (P7). Others emphasized warmth and comfort: *"Playing with the ball to tease the dog was very fun. The puppies were cute, and the tactile sensation felt nice."* (P35). This emphasis on active interaction with the puppies also aligns with the observations in Section 3.4.2, where participants spent longer time in interactive scenes, suggesting that direct interaction with the puppies contributed to both longer engagement in the scene and stronger affective responses.

While both conditions elicited highly positive emotions, the interactive version strengthened these responses by introducing playfulness and tactile engagement, leading to greater immersion, joy, and a stronger sense of control.

3.4.5.3 Topics: Jetty at Lake

The *Jetty at Lake* scene elicited emotions with high valence (M = 6.86 ± 1.42, Med = 7.00), low arousal (M = 4.33 ± 2.41, Med = 4.00), and high dominance (M = 6.67 ± 1.88, Med = 7.00) in the non-interactive condition. Participants frequently associated the scene with calmness and serenity, using terms such as "peaceful" (N = 6), "calm" (N = 4), and "sound" (N = 3). LDA confirmed these impressions, highlighting themes of calmness, restorative nature, and quietness, with occasional mentions of loneliness or ordinariness. Several participants pointed to the natural setting as restorative: *"It felt peaceful and close to nature."* (P55); *"The sound of water made me feel increasingly calm."* (P52).

In the *interactive version*, participants could throw paper airplanes and stones into the lake, which altered the quality of the experience. Under this condition, valence increased (M = 7.48 ± 1.35, Med = 8.00), arousal slightly decreased (M = 4.07 ± 2.42, Med = 4.00), and dominance was rated higher (M = 7.26 ± 1.65, Med = 7.50). As we observed in Figure 3.13, participants frequently mentioned "paper airplane" (N = 9), "stone" (N = 7), "peaceful" (N = 7), "calm" (N = 8), and "relaxed" (N = 4). LDA suggested that while the themes of peace and nature remained central, the addition of interactive elements introduced feelings of fun, relaxation, and control. Participants described them as enjoyable and soothing: *"It felt open, comfortable, and interesting. The paper airplane and throwing stones made me feel calm"* (P6); *"The paper airplane and stones helped me release stress, and the ripples from throwing them felt peaceful"* (P9).

Across conditions, participants consistently associated the *Jetty at Lake* with relaxation. The interactive features deepened this effect by combining the natural scenery with simple actions, making the experience tranquil and engaging. The addition of paper airplanes and stones thus reinforced the calming qualities of the scene while enhancing participants' sense of involvement and control.

3.4.5.4 Topics: Solitary Confinement

The *Solitary Confinement* scene elicited emotions with low valence (M = 3.74 ± 1.89, Med = 4.00), mid arousal (M = 5.17 ± 2.12, Med = 5.00), and low dominance (M = 3.69 ± 1.98, Med = 3.50) in the non-interactive condition. Participants frequently mentioned "oppressive" (N = 6), "sound" (N = 4), "space" (N = 3), "small" (N = 3), and "dripping water" (N = 3). LDA revealed themes of small space, disturbing sounds, and the prison setting. Participants often described the space as narrow and uncomfortable: *"It felt oppressive, and the room was small."* (P53). Others pointed to the audio design

as particularly disturbing: *"The dripping water sound was loud and annoying, making me feel anxious."* (P79).

In the *interactive condition*, where participants could interact with the cup, book, and door, the scene elicited slightly higher valence (M = 4.52 ± 2.02, Med = 5.00), similar arousal (M = 5.10 ± 2.09, Med = 5.00), and higher dominance (M = 4.86 ± 2.53, Med = 5.00). Frequently mentioned words included "space" (N = 13), "small" (N = 11), "sound" (N = 10), "dripping water" (N = 5), and "oppressive" (N = 3). LDA indicated that although participants still described the space as confined and unpleasant, the possibility of interacting with the environment introduced a sense of relief or playfulness. For example, one participant noted: *"It felt restrictive, the space was very small, and the dripping water sound made me feel tense and worried."* (P36). Others highlighted playful aspects of interaction: *"Throwing things at the door and hearing the sound felt provocative and satisfying."* (P7).

Overall, while both conditions emphasized the oppressive qualities of the space and the disturbing soundscape, the interactive version allowed participants to vent their feelings and exert some control, making the experience slightly less negative.

3.4.5.5 Topics: Shouting Man with Gun

The *Shouting Man with Gun* scene in non-interactive condition aimed to elicit emotions with low valence (M = 4.33 ± 1.86, Med = 5.00), high arousal (M = 6.81 ± 1.61, Med = 7.00), and low dominance (M = 4.12 ± 1.98, Med = 4.00). Frequently mentioned words included "man" (N = 14), "sudden" (N = 14), "shout" (N = 13), "frightening" (N = 6), and "gun" (N = 4). LDA emphasized themes of sudden appearance, shouting, and danger, which reinforced feelings of fear and surprise. Participants described the scene as startling and frightening: *"The man and his sudden shout were frightening"* (P63); *"The view outside the window was calm, but the sudden entrance of the man was frightening"* (P72).

In the *interactive condition*, ratings with higher valence (M = 4.93 ± 2.15, Med = 5.00), identical arousal (M = 6.83 ± 2.14, Med = 7.50), and higher dominance (M = 4.50 ± 2.44, Med = 4.00). Participants frequently mentioned "shield" (N = 15), "man" (N = 13), "safety" (N = 8), "sudden" (N = 5), and "gun" (N = 4). LDA indicated that, while the man's sudden entrance and shouting continued to evoke fear, the shield was perceived as protection that reduced helplessness. As one participant explained: *"In the enclosed space, the man shouted, and his face and the way he pointed the gun at me made me feel scared. The shield, in contrast, gave me a sense of safety and immersion"* (P12). Another described:

"The sudden event was unexpected and very frightening, but the shield gave me a slight sense of safety" (P33).

In summary, the *Shouting Man with Gun* scene elicited high-arousal, negative valence, dominated by fear and surprise. While both conditions produced similar responses, the interactive shield introduced a partial sense of safety, restoring a degree of agency.

3.4.5.6 Topics: Surrounded by Elephants

In the non-interactive condition, the *Surrounded by Elephants* scene was rated with high valence (M = 6.90 ± 1.76, Med = 7.00), high arousal (M = 5.76 ± 2.21, Med = 6.00), and mid dominance (M = 5.81 ± 2.30, Med = 6.00). Participants frequently mentioned "elephant" (N = 15), "scared" (N = 6), "large" (N = 4), "grassland" (N = 3), and "comfortable" (N = 2). LDA identified themes of openness, realism, and fear. One participant highlighted the vast natural setting: *"The overall scene had an open view that matched my expectations."* (P51). Others described mixed feelings: *"It felt realistic, and it seemed like the elephant was going to attack me. I felt scared, fearful, and tense."* (P57).

With interaction, ratings remained highly positive, with higher valence (M = 7.07 ± 1.44, Med = 7.00), slightly lower arousal (M = 5.71 ± 2.43, Med = 6.50), and higher dominance (M = 6.48 ± 2.03, Med = 7.00). Participants often mentioned "elephant" (N = 17), "grassland" (N = 8), "banana" (N = 4), "large" (N = 4), and "relaxed" (N = 3). LDA revealed themes of awe and nervousness, but also noted how feeding and touching the elephants provided moments of relief and enjoyment. As one participant observed: *"The huge elephants created a slight sense of pressure and made me feel a bit nervous. The banana feeding interaction gave me a sense of awareness and made my mood feel calmer."* (P9). P17 echoed this mix of fear and enjoyment: *"The elephants were a bit scary because they were so big, but watching them eat bananas made me feel a bit more relaxed."*

To summarize, the *Surrounded by Elephants* scene elicited high-valence and high-arousal, with participants describing relaxation alongside moments of fear. According to our results, in the interactive condition, feeding and touching the elephants were often described as calming and enjoyable, relieving tension and fostering fun and involvement.

3.4.5.7 Overview of topic modeling

In summary, across all six scenes, the virtual environments effectively elicited the intended emotions, and our interactions with them further shaped participants' responses. In the **non-interactive condition**, participants emphasized environmental atmosphere and passive reception, often reflecting a lack of control. In contrast, **the interactive condition** foregrounded agency and engagement. In negative scenes (e.g., *Solitary Confinement* and *Shouting Man with Gun*), our interaction designs provided avenues for coping and relief, whereas in positive or neutral scenes (e.g., *Puppies* and *Jetty at Lake*), they enhanced enjoyment and immersion. Overall, these results show that our interaction with the scenes did not simply intensify emotions but modulated them in context, highlighting the effectiveness of scene-tailored interaction in creating rich emotional experiences in VR.

CHAPTER 4

An LLM-Assisted Toolkit for Inspectable Multimodal Emotion Data Annotation

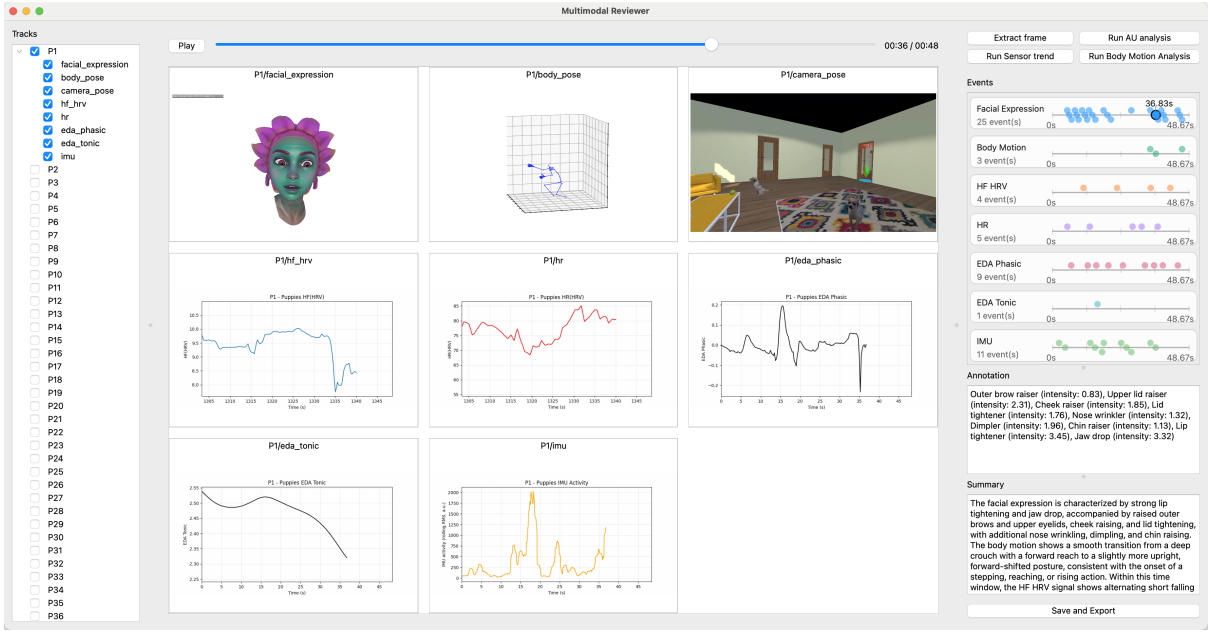


FIGURE 4.1. Toolkit interface for multimodal emotion data inspection and annotation.

Building on the interactive VR study in Chapter 3, this chapter turns to the inspection and annotation of the multimodal emotion data generated during these VR experiences. The previous study [58] collected self-reports, facial expression videos, first-person scene videos, body motion, and physiological signals using VR headsets and wearable sensors. These signals are related to emotional experience, but they are heterogeneous, asynchronous, and distributed across different temporal scales, making annotation laborious, expensive, and time-consuming. Prior work suggests that LLMs can support data inspection and candidate annotation while preserving human verification and analyst control [52, 41, 122]. To address this challenge, this chapter introduces an LLM-assisted toolkit for aligned visualization, event-level evidence retrieval, and inspectable multimodal emotion annotation.

4.1 Proposed Approach

We present an LLM-assisted toolkit that operationalizes a three stage data handling workflow for multimodal emotion data annotation. The toolkit transforms heterogeneous recordings into inspectable, event-centered outputs that support visual inspection of signal changes, emotion-related event detection and keyframe retrieval, and evidence-grounded annotations. It supports data modalities including but not limited to video streams and structured time-series signals.

We demonstrate the toolkit on our multimodal VR emotion recordings from 84 participants across six emotion elicitation scenes, including avatar-based facial expression video, first-person view video, structured body motion, blood volume pulse (BVP), heart rate (HR), electrodermal activity (EDA), and inertial measurement unit (IMU) signals. As shown in Figure 4.1, the interface provides synchronized multimodal tracks, event timelines, and annotation views for event-centered inspection and labeling.

4.1.1 Preprocess and visualization

In the first stage, the toolkit performs preprocessing and unified visualization. The goal is to make heterogeneous modalities jointly inspectable on a shared timeline so that analysts can efficiently observe signal changes and localize potential events. We first standardize each modality with consistent session metadata, including participant identifiers, VR scene identifiers, sampling rates, and timestamps. For structured signals, we render each stream into a video track, so that time series can be inspected with the same interaction paradigm as videos. This enables time locked playback and side by side comparison across different modality tracks in a single interface.

We then synchronize all tracks to a unified time axis and persist the alignment as an index, which supports navigation, zooming, and reproducible retrieval of any time window across modalities. The resulting UI organizes sessions by participant and VR scene, providing an overview for rapid scanning and detailed views for inspecting fine grained temporal patterns behavioral and physiological signal peaks. This stage outputs aligned visual tracks and a session level index that serve as the basis for event mining and subsequent annotation.

4.1.2 Event detection and retrieval

In the second stage, the toolkit detects signal change events and retrieves the corresponding keyframes and time windows across modalities. For facial expression videos, we perform Action Units (AUs) [28] peak detection using OpenFace [5], and treat each peak as a candidate event. For body motion signals, we compute motion energy measures to identify motion events [96]. For physiological signals, we extract peak and trend change windows to capture candidate events. All candidates are aligned to the unified time axis and aggregated into an event layer in the UI, enabling analysts to jump to events, compare cross-modal context, and refine event boundaries. We render this layer as an interactive event timeline, where clicking an event marker seeks to its timestamp and shows the aligned frames across modalities. Analysts can verify, edit, or discard candidate events in the UI. The toolkit then packages each event with traceable pointers to the original data, including keyframes and time windows for all modalities.

4.1.3 LLM annotations

In the third stage, the toolkit supports LLM-assisted emotional annotation for extracted events. The goal of this stage is to convert each event window into standardized annotations and evidence summaries that can be used for emotion related event detection and for constructing training data for downstream emotion recognition models.

We designed a set of modality specific preprocessing tools and prompt templates to guide the LLM to produce a structured annotation with consistent fields. We employ GPT-5.2 as the backend LLM to leverage its multimodal reasoning capabilities. For facial expression, we map peak AU frames to short textual descriptions [19]. For body motion, the LLM generates annotations that describe the skeleton sequence within the event window using concise posture and movement cues [80]. For physiological signals, the LLM summarizes the window using trend and peak descriptors, including rising or falling segments, local extrema, and duration [70]. For the first person view stream, the LLM captures contextual information such as activities and environment cues that support interpretation of the event. Finally, the LLM refines the annotations by aggregating unimodal descriptions into a detailed multimodal description for each event [19], which is added to the event packet as an additional emotional descriptor.

Analysts review the generated annotation in the UI, jump to the referenced time ranges for verification, and edit or discard incorrect fields. Finalized annotations are then exported as structured records for downstream use.

CHAPTER 5

MoE-MER: A Mixture-of-Experts LLM Framework for Multimodal Emotion Recognition in Virtual Reality

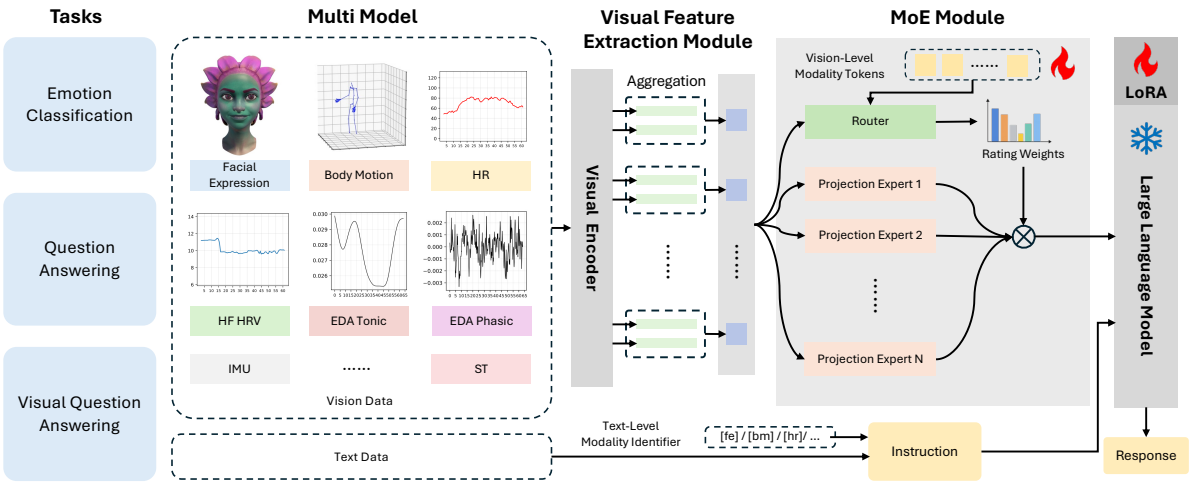


FIGURE 5.1. Overall architecture of MoE-MER, which consists of a universal vision feature extraction module, a MoE module and an LLM. MoE-MER can perform three different emotion tasks including emotion classification, question answering and visual question answering

Building on the multimodal data annotation workflow introduced in Chapter 4, this chapter moves from data inspection and annotation to multimodal emotion recognition. The previous chapter focused on organizing heterogeneous VR emotion data into aligned and inspectable event-level evidence. However, a further challenge is how such heterogeneous evidence can be used for emotion recognition while supporting human-understandable explanations. To address this challenge, this chapter proposes an LLM-based Mixture-of-Experts framework for multimodal emotion recognition, which integrates visual, physiological, and contextual cues for prediction and explanation.

5.1 Methodology

With the primary goal of achieving a unified multimodal emotion recognition model and mitigating task interference in multi-task learning, we design the overall architecture of MoE-MER as illustrated in Figure 5.1. It contains three components: a universal vision feature extraction module, a modality-aware MoE module, and an LLM. Detailed descriptions are presented in the following sections.

5.1.1 Visual feature extraction module

MoE-MER processes heterogeneous modalities through a unified visual pathway. For each sample, the input may consist of a single image or a set of images converted from different modalities, such as facial expression representations, body motion visualizations, and physiological waveform plots. A shared visual encoder is applied to each image to produce a sequence of visual tokens that capture informative local patterns and global context in a common embedding space.

When the input contains multiple images, or when a single image is high-resolution, the visual encoder can produce a long token sequence, which increases the computational cost in both training and inference. To improve efficiency, we apply a lightweight token aggregation step after the visual encoder to compress the visual tokens into a compact sequence with a controlled length. Concretely, we merge local groups of adjacent tokens within each image and project them into aggregated tokens, producing a shorter representation that preserves key visual evidence. The resulting aggregated tokens provide a compact visual context for the downstream MoE module and the LLM, while retaining modality source information for subsequent modality-aware processing.

5.1.2 MoE module

After visual feature extraction, MoE-MER uses a modality-aware MoE module to adapt representations from different modalities before they are fed into the LLM. Although heterogeneous signals are converted into visual inputs, their characteristics still differ across modalities, including facial images, skeleton motion plots, and physiological waveform charts. If a single shared projection is used for all modalities, modality-specific cues can be weakened, and negative transfer may occur because modalities follow different data statistics. We therefore introduce modality experts that perform modality-specific transformations while keeping a unified interface for downstream generation.

The MoE module contains multiple projection experts and a lightweight router. Each expert maps visual tokens from one modality, or from a small set of closely related modalities, into the embedding space used by the LLM. For modality-aware routing, each input image and its tokens are paired with an explicit modality tag attached to the token sequence. Conditioned on the aggregated visual tokens and the modality tag, the router outputs mixture weights over experts. The projected visual representation is then computed as a weighted combination of expert outputs, allowing tokens from different modalities to be transformed by different experts.

Finally, the projected tokens from all modalities are merged into a shared visual context and provided to the LLM together with the textual instruction. This design supports unified generation across emotion classification, question answering, and visual question answering.

5.1.3 Large language model

The MoE module focuses on modality differences, while the task is specified by the language instruction. MoE-MER uses an LLM as a unified interface for multi-task outputs. The LLM receives a shared visual context derived from multiple modalities and a text instruction that specifies what to predict. The visual context is formed by merging the projected tokens produced by the MoE module across all available modalities in the sample. This design allows the LLM to attend to facial, motion, and physiological signals within a single sequence, while keeping the downstream reasoning and generation in one model.

To support different tasks, we rely on an instruction-driven prompting format. The instruction explicitly indicates the task type and expected output, such as predicting an emotion label for emotion classification, answering a textual question for question answering, or answering a question grounded in the provided visual evidence for visual question answering. In addition, a short task tag is prepended to the instruction to standardize the input format across tasks. Importantly, this task specification is handled at the language level and does not change the modality-aware routing in the MoE module.

Given the visual context and instruction tokens, the LLM generates the response autoregressively. To adapt the base LLM to our domain and prompting format efficiently, we use low-rank adaptation (LoRA) for fine-tuning. LoRA is applied to the main linear projections of the transformer, enabling parameter-efficient training while keeping most pretrained weights fixed.

CHAPTER 6

Discussion

Next, we answer our research questions, position our work in existing work, and lay out its implications and limitations.

6.1 RQ1: Does the added scene *Surrounded by Elephants* reliably and effectively elicit High Arousal and High Valence emotions?

We extended the emotion elicitation dataset [46] by adding a new VR scene, *Surrounded by Elephants*, filling the gap in the HAHV quadrant. We validated this new VR scene through a user study, following the approach by Jiang et al. [46]. Our findings reveal that compared to 360° video, the VR scene elicited emotions with significantly higher valence and dominance. The higher valence ratings may suggest that participants experienced more positive emotions in the VR condition, consistent with prior research on positive emotion elicitation, which shows that an enhanced sense of presence in VR can strengthen affective responses [97]. The higher dominance ratings further indicate that participants felt more engaged and were able to actively shape and control their experience through self-initiated behaviors in the VR environment, aligning with studies linking presence and control in VR to greater perceived dominance [102, 111]. In contrast, the difference in arousal was not significant, indicating that we found no difference between the VR scene and the 360° video along this dimension.

Overall, our scene demonstrated that the *Surrounded by Elephants* VR scene can reliably and effectively elicit HAHV emotions, due to its immersive nature, as participants were able to *e.g.*, approach or observe the elephants.

6.2 RQ2: How does object-level interaction influence subjective and physiological measures of emotional response compared to a non-interactive baseline?

We further extended the emotion elicitation dataset [46] by adding interaction to the scenes. Our findings demonstrate that, relative to the non-interactive baseline, *object-level interaction* did not exhibit a uniform main effect on self-reported valence, arousal, or dominance, with differences appearing primarily as scene-dependent effects. However, the interaction patterns indicated that affective interaction design consistently elevated Dominance in specific contexts by allowing participants to feel more in control of the emotional experience rather than passively receiving it. This suggests that perceived control is one mechanism through which object-level interaction shaped emotional experience, as interactive objects gave participants concrete ways to act within and respond to the scene.

The interactive versions showed longer engagement time and broader spatial exploration, indicating that participants were more willing to stay and explore when given interaction options. This extended engagement provided more opportunities for emotional responses, reflected in higher valence and dominance ratings, as well as in physiological patterns of higher HF HRV and lower HR.

Furthermore, our findings align with prior work and show that *object-level interaction* transforms passive experience into goal-directed exploration [21]. Interactive objects provide clear action goals and feedback, prompting participants to actively perceive the virtual environment. Once participants know that interactive objects exist in the scene, they are more willing to remain within the scene and explore it. In low-valence or tense scenes, such as *Tunnel*, *Solitary Confinement*, and *Shouting Man with Gun*, instrumental or defensive interaction offers ways to cope and regulate, reducing the tendency to exit quickly due to discomfort. Similarly, in scenes with high valence, such as *Puppies*, *Jetty at Lake*, and *Surrounded by Elephants*, interaction seems to sustain engagement and enjoyment, thereby extending participation.

Similarly, our findings demonstrate that *object-level interaction* changes participants' spatial exploration patterns in virtual scenes. Compared with the non-interactive version, spatial engagement shifts from simple surveying to goal-directed engagement with interactive objects. We observe that even when some objects are not configured as interactive, participants still attempt to engage with them, concentrating movement around potential interactive hotspots within a limited activity area. In the *Puppies* scene,

because the positions of the ball and the puppies can be changed, participants show broader spatial engagement with the environment. The playful and repeatable feedback from petting the puppies and playing fetch with the ball further sustains this goal-directed engagement, helping explain the particularly long engagement time in the interactive condition. This goal-directed engagement has been shown to relate to emotional responses [33]: at the subjective level, it increases dominance, while the direction of valence and arousal depends on the affective character of the scene [90]; at the physiological level, it aligns with the overall pattern of higher HF HRV and lower HR [57], and increased SCR count and amplitude [23]. Future work could extend this analysis by examining transient SCR fluctuations around key interaction events to capture finer-grained dynamics of physiological arousal [13].

Overall, our findings suggest that interaction does not merely enhance emotional elicitation; by providing concrete goals, predictable feedback, and visible information, it reduces uncertainty and shapes the emotional experience in a controllable, context-sensitive way, rather than relying on passive exposure alone.

6.3 RQ3: What is the relationship between subjective self-reports and physiological arousal in response to affective interactions in VR?

In the interactive condition, across all scenes, we observed changes in physiological arousal that reflect two complementary dimensions. Higher HF HRV and lower HR indicate lower tonic arousal intensity, pointing to reduced physiological arousal [3, 57], while higher SCR count and amplitude reflected more frequent and stronger phasic activations, pointing to momentary sympathetic responses, e.g., HR accelerations, SCR responses [23, 13]. These physiological patterns consistently accompanied higher subjective valence and dominance, suggesting that participants felt more positive and in control, allowing them to maintain overall physiological relaxation interspersed with moments of high reactivity to interactive stimuli.

The relationship between subjective and physiological arousal showed dependence on the scene. For example, in the LALV scene *Tunnel*, interaction reduced tension and uncertainty, leading to lower subjective arousal consistent with greater predictability and control. In *Solitary Confinement*, interaction introduced executable goals and immediate feedback, shifting experience from passive low activation to object-directed engagement; subjective arousal and valence increased, while physiological measures

indicated that this activation remained controllable rather than stressful. In the LAHV scenes, the effect depended on context: relaxation-oriented interaction reduced subjective arousal in *Jetty at Lake*, whereas play-oriented interaction increased it in *Puppies*. For high-arousal scenes, *Surrounded by Elephants* (HAHV) showed only a modest decrease in subjective arousal, while *Shouting Man with Gun* (HALV) yielded increased arousal despite weaker increases in HF HRV, suggesting vigilance combined with emerging control.

In summary, interaction was associated with a consistent increase in valence and dominance, while physiological arousal showed a dual profile of tonic relaxation and phasic activation [3, 57, 23]. Subjective arousal, in contrast, was more malleable, varying with scene context: decreasing in relaxation-oriented scenarios and rising under threat or active engagement [78, 119, 81]. This partial dissociation between subjective and physiological arousal aligns with broader evidence that self-reports are shaped by cognitive appraisal of controllability and meaning [36], whereas physiological measures capture the temporal dynamics of tonic versus phasic responses [4], which may not always be directly mapped onto subjective experience [87, 86].

6.4 RQ4: How can an LLM-assisted toolkit support inspectable and event-centered multimodal emotion data annotation?

We developed an LLM-assisted toolkit that bridges our multimodal emotion data collected from VR experience and downstream MER modeling by operationalizing an inspectable, event-centered annotation workflow. The toolkit first preprocesses and aligns heterogeneous modalities on a shared timeline, rendering structured time-series signals as video tracks to support cross-modal consistency checks. It then detects candidate signal-change events and packages each event into an event packet with traceable pointers to aligned keyframes and time windows across modalities. Finally, an LLM generates structured, modality-specific annotations using tools and prompt templates, while analysts verify and edit the outputs in the interface, enabling more scalable construction of training-ready annotations.

6.5 RQ5: How can an LLM-based MoE framework improve multimodal fusion and provide human-understandable explanations for MER?

We designed MoE-MER as an LLM-based MoE framework to support multimodal fusion and unified multi-task outputs for MER. MoE-MER processes heterogeneous modalities through a unified visual pathway: each modality is converted into a single image or a set of images, a shared visual encoder extracts visual tokens, and a lightweight token aggregation step compresses long token sequences for efficient training and inference. It then uses a modality-aware MoE module, where projection experts transform modality-tagged tokens into the LLM embedding space and a router produces mixture weights conditioned on the aggregated tokens and modality tags, aiming to preserve modality-specific cues and mitigate negative transfer while keeping a unified interface. Finally, the LLM takes the merged projected tokens as a shared visual context together with an instruction that specifies the task (i.e., emotion classification, question answering, and visual question answering), enabling consistent outputs in a human-readable form. Since we currently present a method design, our contribution here is the framework and its intended mechanisms for modality-adaptive fusion and instruction-driven, explanation-oriented outputs.

6.6 Positioning within Existing Literature

Compared to previous studies using the original 360° videos [65, 108] and VR scenes [46], our findings were broadly consistent with prior work but revealed some systematic deviations. As shown in Figure 3.5, LALV and HALV scenes elicited higher valence at comparable arousal levels, while HAHV and LAHV scenes elicited higher arousal at comparable valence. These deviations may be explained by methodological factors. Jiang et al. [46] conducted their study with Prolific participants completing VR experiences in different environments, where experimental control is limited, and variability can undermine consistency [93]. By contrast, our controlled laboratory setting ensured standardized equipment and procedures, conditions shown to enhance immersion and presence in VR [25, 112]. To summarize, our findings align with prior work [46], but reveal systematic deviations in emotional responses, likely driven by the enhanced experimental control and immersion afforded by the laboratory setting compared to less controlled environments.

6.7 Implications for Research and Design

Our findings suggest that interaction in VR should be reconsidered not only as a method for eliciting emotions but also as a medium for modulating them. Specifically, we observed that interaction amplified positive affect in playful contexts, e.g., *Puppies* and *Surrounded by Elephants*, while reducing perceived threat and enhancing dominance in stressful scenes, e.g., *Tunnel* or *Solitary Confinement*. This indicates that the agency has the potential to function as a mechanism for emotional regulation in VR, reinforcing that immersive systems can shape coping and affective trajectories rather than simply intensifying responses [103, 76]. To facilitate such research, our validated dataset serves as a base for the HCI community to investigate how different object-level interactions influence emotional experiences in VR. Beyond the dataset, our findings offer a theoretical foundation for mental health research, specifically supporting a shift from passive exposure to active coping paradigms where affective interaction fosters a sense of control and safety [9, 114]. Consequently, future work should extend beyond the valence–arousal model and develop measures that capture coping, relief, and sustained immersion. Such perspectives would position VR not only as an affective stimulus, but as a research tool for understanding and supporting emotion regulation across domains such as mental health [85], training [53], and education [84].

For design practice, our results highlight the importance of tailoring both scene content and interaction to the intended affective experience. Designing a new affective VR scene should not only involve selecting content that matches a target affective quadrant, but also considering how the scene produces that emotion in a safe and coherent way. Specifically, visual elements, auditory cues, object affordances, and interaction feedback jointly shape the emotional meaning of the scene. For example, in Surrounded by Elephants, the open natural environment supported a positive emotional meaning, while the elephants' trumpeting sounds, proximity, scale, and movement contributed to arousal; feeding or touching the elephants made this encounter actionable through interaction. These elements should fit users' understanding of the situation and be calibrated to avoid distress or cybersickness. Interaction should also follow the emotional role of the scene and support the intended affective experience. In threatening contexts, giving users agency to act (*e.g.*, controlling light in a dark tunnel) may alleviate distress and restore balance, while in positive contexts, playful or exploratory actions can heighten enjoyment and sustain engagement. These insights can be extended to diverse application areas. In VR game design, developers can adaptively adjust game difficulty or mechanics based on players' emotional state, providing empowering tools when anxiety is too high, or playful interactive objects when engagement drops [31].

In social VR, designing object-level interaction can provide shared focus and foster group collaboration and connection [107]. In VR learning, playful interaction designs can promote active engagement and boost the sense of dominance, thereby enhancing learning outcomes [107]. However, effective design requires careful calibration. Prior studies on emotion elicitation in VR caution that poorly calibrated interaction risks overstimulation or heightened anxiety [103], a concern our results corroborate. Thus, effective interaction design in VR requires balancing environmental atmosphere with user agency, ensuring that interaction supports adaptive regulation and meaningful emotional experiences.

6.8 Limitations and Future Work

Our thesis has several limitations that should be considered when interpreting the findings across the full workflow. First, our interaction study focused on object-level interactions; more complex forms, such as social interaction or dynamic environmental feedback, were not within our scope. Some implemented actions may also resemble game mechanics, which could influence how users attribute and regulate their emotional responses across groups. In addition, the study measured short-term responses to each VR scene and did not examine whether the increased engagement and exploration observed in interactive scenes would persist after repeated exposure. Second, our measures relied on self-report (SAM) and basic physiological signals (EDA and BVP). While useful, these indicators are indirect and subject to substantial individual variability, limiting fine-grained interpretations of the regulation processes we observed. In addition, although our toolkit operationalizes an inspectable, event-centered workflow for alignment, candidate event retrieval, and LLM generation with analyst verification, we have not yet evaluated its impact on inspection efficiency and annotation outcomes against standard practices. Finally, we present MoE-MER as a method design for modality-aware fusion and unified multi-task outputs, but it does not yet include experimental results; therefore, this thesis does not claim improvements in recognition performance or explanation quality.

Future work can extend this thesis along several lines. For interaction research, we will broaden both the forms of interaction and the methods of assessment, and examine whether interaction effects on engagement, dominance, and emotional response persist across repeated sessions. Recent surveys highlight that multimodal affect recognition, integrating physiology with facial expressions, voice, and eye-tracking, provides a richer picture of emotional dynamics than single-modality approaches [1]. Similarly, researchers argue that affective touch and embodied sensing are key to advancing emotion research in immersive contexts [141]. These perspectives point beyond the valence–arousal–dominance model toward

frameworks that capture coping, relief, and sustained engagement. Incorporating multimodal elicitation, including haptic feedback, may therefore enable more embodied emotional experiences in VR and deeper insights into how interaction might shape emotion elicitation and emotion regulation. For the toolkit, we will conduct a user study with domain experts to compare inspection efficiency and annotation outcomes against standard practices, and validate the workflow on additional datasets with more modalities and adaptable annotation schemas. For modeling, we will complete a systematic evaluation of MoE-MER to determine its effectiveness for multimodal fusion and unified instruction-driven outputs using our dataset. We will also assess whether generated outputs remain grounded in multimodal evidence by analysing routing patterns and verifying responses under modality corruption or removal.

CHAPTER 7

Conclusion

This thesis investigated how object-level interaction in VR modulates emotion elicitation, and how LLMs can support inspectable annotation and multimodal emotion recognition using multimodal emotional data collected during VR. To address gaps in both affective coverage and interaction design, we extended an existing VR emotion elicitation dataset by adding a new High-Arousal High-Valence scene, Surrounded by Elephants, and by implementing interactive and non-interactive versions of each scene to isolate the effect of object-level interaction under controlled conditions. Across studies, we combined self-reported affect with physiological sensing to capture both post-hoc appraisal and time-varying autonomic responses, showing that interaction shapes emotional experience in a context-sensitive way rather than producing uniform changes.

To support downstream analysis and modeling, we developed an LLM-assisted, event-centered toolkit that aligns heterogeneous recordings, retrieves candidate events as traceable event packets, and drafts structured annotations under human verification and editing. Building on these event-level outputs, we proposed MoE-MER as an LLM-based MoE framework designed for modality-adaptive fusion and instruction-driven outputs, with the goal of producing emotion predictions and human-understandable, evidence-linked responses. Together, these contributions form a systematic pipeline from controlled interactive VR emotion elicitation to inspectable annotation and LLM-based multimodal emotion recognition.

Bibliography

[1] Sharmeen M Saleem Abdullah Abdullah, Siddeeq Y Ameen Ameen, Mohammed AM Sadeeq, and Subhi Zeebaree. 2021. Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends*, 2(01):73–79.

[2] Mostafa M Amin, Rui Mao, Erik Cambria, and Björn W Schuller. 2024. A wide evaluation of chatgpt on affective computing tasks. *IEEE Transactions on Affective Computing*, 15(4):2204–2212.

[3] Bradley M Appelhans and Linda J Luecken. 2006. Heart rate variability as an index of regulated emotional responding. *Review of general psychology*, 10(3):229–240.

[4] Ebrahim Babaei, Benjamin Tag, Tilman Dingler, and Eduardo Velloso. 2021. A critique of electrodermal activity practices at chi. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

[5] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *WACV*, pages 1–10.

[6] Soumya C Barathi, Michael Proulx, Eamonn O'Neill, and Christof Lutteroth. 2020. Affect recognition using psychophysiological correlates in high intensity vr exergaming. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

[7] Lyn Bartram, Abhisekh Patra, and Maureen Stone. 2017. Affective color in visualization. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 1364–1374.

[8] Allison Bayro and Heejin Jeong. 2025. A systematic review of experimental protocols: Towards a uniform framework in virtual reality affective research. *IEEE Transactions on Affective Computing*.

[9] Imogen H Bell, Jennifer Nicholas, Mario Alvarez-Jimenez, Andrew Thompson, and Lucia Valmaggia. 2020. Virtual reality as a clinical tool in mental health research and practice. *Dialogues in clinical neuroscience*, 22(2):169–177.

[10] Stuart M Bender and Billy Sung. 2021. Fright, attention, and joy while killing zombies in virtual reality: A psychophysiological analysis of vr user experience. *Psychology & Marketing*, 38(6):937–947.

[11] Alberto Betella and Paul FMJ Verschure. 2016. The affective slider: A digital self-assessment scale for the measurement of human emotions. *PloS one*, 11(2):e0148037.

[12] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[13] Wolfram Boucsein. 2012. *Electrodermal activity*. Springer science & business media.

[14] Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59.

[15] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

[16] Polona Caserman, Augusto Garcia-Agundez, Alvar Gámez Zerban, and Stefan Göbel. 2021. Cybersickness in current-generation virtual reality head-mounted displays: systematic review and outlook. *Virtual Reality*, 25(4):1153–1170.

[17] Feiyu Chen, Jie Shao, Shuyuan Zhu, and Heng Tao Shen. 2023. Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10761–10770.

[18] Ke Chen, Lei Xu, and Huisheng Chi. 1999. Improved learning algorithms for mixture of experts in multiclass classification. *Neural networks*, 12(9):1229–1252.

[19] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. 2024. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37:110805–110853.

[20] Alice Chirico, Francesco Ferrise, Lorenzo Cordella, and Andrea Gaggioli. 2018. Designing awe in virtual reality: An experimental study. *Frontiers in psychology*, 8:2351.

[21] Athanasios Christopoulos, Marc Conrad, and Mitul Shukla. 2018. Increasing student engagement through virtual interactions: How? *Virtual Reality*, 22(4):353–369.

[22] John R Crawford and Julie D Henry. 2004. The positive and negative affect schedule (panas): Construct validity, measurement properties and normative data in a large non-clinical sample. *British journal of clinical psychology*, 43(3):245–265.

[23] Hugo D Critchley. 2002. Electrodermal responses: what happens in the brain. *The Neuroscientist*, 8(2):132–142.

[24] Mihaly Csikszentmihalyi and Reed Larson. 1987. Validity and reliability of the experience-sampling method. *The Journal of nervous and mental disease*, 175(9):526–536.

[25] James J Cummings and Jeremy N Bailenson. 2016. How immersive is enough? a meta-analysis of the effect of immersive technology on user presence. *Media psychology*, 19(2):272–309.

[26] Nicoló Dozio, Federica Marcolin, Giulia Wally Scurati, Luca Ulrich, Francesca Nonis, Enrico Vezzetti, Gabriele Marsocci, Alba La Rosa, and Francesco Ferrise. 2022. A design methodology for affective virtual reality. *International journal of human-computer studies*, 162:102791.

[27] Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

[28] Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

[29] Aviv Elor, Asiiah Song, and Sri Kurniawan. 2021. Understanding emotional expression with haptic feedback vest patterns and immersive virtual reality. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 183–188. IEEE.

[30] Sergio Estupiñán, Francisco Rebelo, Paulo Noriega, Carlos Ferreira, and Emília Duarte. 2014. Can virtual reality increase emotional responses (arousal and valence)? a pilot study. In *International conference of design, user experience, and usability*, pages 541–549. Springer.

[31] Stephen H Fairclough. 2009. Fundamentals of physiological computing. *Interacting with computers*, 21(1-2):133–145.

[32] Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. 2025. Emoe: Modality-specific enhanced dynamic emotion experts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14314–14324.

[33] Yu-Min Fang. 2024. Exploring usability, emotional responses, flow experience, and technology acceptance in vr: a comparative analysis of freeform creativity and goal-directed training. *Applied Sciences*, 14(15):6737.

[34] Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*.

[35] Marco Granato, Davide Gadia, Dario Maggiorini, and Laura A Ripamonti. 2020. An empirical study of players' emotions in vr racing games based on a dataset of physiological data. *Multimedia tools and applications*, 79(45):33657–33686.

[36] James J Gross. 2015. Emotion regulation: Current status and future prospects. *Psychological inquiry*, 26(1):1–26.

[37] Kunal Gupta, Sam WT Chan, Yun Suen Pai, Nicholas Strachan, John Su, Alexander Sumich, Suranga Nanayakkara, and Mark Billinghurst. 2022. Total vrecall: Using biosignals to recognize emotional autobiographical memory in virtual reality. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(2):1–21.

[38] Kunal Gupta, Yuewei Zhang, Tamil Selvan Gunasekaran, Nanditha Krishna, Yun Suen Pai, and Mark Billinghurst. 2024. Caevr: Biosignals-driven context-aware empathy in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 30(5):2671–2681.

[39] Martin Halvey, Michael Henderson, Stephen A Brewster, Graham Wilson, and Stephen A Hughes. 2012. Augmenting media with thermal stimulation. In *International Conference on Haptic and Audio Interaction Design*, pages 91–100. Springer.

[40] Zhaopei Huang, Jinming Zhao, and Qin Jin. 2024. Ecr-chain: Advancing generative language models to better emotion-cause reasoners through reasoning chains. *arXiv preprint arXiv:2405.10860*.

[41] Maeve Hutchinson, Radu Jianu, Aidan Slingsby, and Pranava Madhyastha. 2024. Llm-assisted visual analytics: Opportunities and challenges. *arXiv:2409.02691*.

[42] Syem Ishaque, Alice Rueda, Binh Nguyen, Naimul Khan, and Sridhar Krishnan. 2020. Physiological signal analysis and classification of stress from virtual reality video game. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society*

(EMBC), pages 867–870. IEEE.

[43] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

[44] Giulio Jacucci, Andrea Bellucci, Imtiaj Ahmed, Ville Harjunen, Michiel Spape, and Niklas Ravaja. 2024. Haptics in social interaction with agents and avatars in virtual reality: a systematic review. *Virtual Reality*, 28(4):170.

[45] Jason Jerald. 2015. *The VR book: Human-centered design for virtual reality*. Morgan & Claypool.

[46] Weiwei Jiang, Maximiliane Windl, Benjamin Tag, Zhanna Sarsenbayeva, and Sven Mayer. 2024. An immersive and interactive vr dataset to elicit emotions. *IEEE Transactions on Visualization and Computer Graphics*.

[47] Crescent Jicol, Chun Hin Wan, Benjamin Doling, Caitlin H Illingworth, Jinha Yoon, Charlotte Headey, Christof Lutteroth, Michael J Proulx, Karin Petrini, and Eamonn O'Neill. 2021. Effects of emotion and agency on presence in virtual reality. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–13.

[48] Xin Jing, Jiadong Wang, Iosif Tsangko, Andreas Triantafyllopoulos, and Björn W Schuller. 2025. Melt: Towards automated multimodal emotion data annotation by leveraging llm embedded knowledge. *arXiv:2505.24493*.

[49] Sarah Jones. 2017. Disrupting the narrative: immersive journalism in virtual reality. *Journal of media practice*, 18(2-3):171–185.

[50] Michael I Jordan and Robert A Jacobs. 1994. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214.

[51] Joohee Jun, Myeongul Jung, So-Yeon Kim, and Kwanguk Kim. 2018. Full-body ownership illusion can change our emotion. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–11.

[52] Hong Jin Kang, Muhammad Ali Gulzar, Nanyun Peng, Miryung Kim, et al. 2024. Human-in-the-loop synthetic text data inspection with provenance tracking. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3118–3129.

[53] Margaret E Kemeny, Carol Foltz, James F Cavanagh, Margaret Cullen, Janine Giese-Davis, Patricia Jennings, Erika L Rosenberg, Omri Gillath, Phillip R Shaver, B Alan Wallace, et al. 2012. Contemplative/emotion training reduces negative emotional behavior and promotes prosocial responses. *Emotion*, 12(2):338.

[54] Angelika C Kern, Wolfgang Ellermeier, and Lina Jost. 2020. The influence of mood induction by music or a soundscape on presence and emotions in a virtual reality park scenario. In *Proceedings of the 15th International Audio Mostly Conference*, pages 233–236.

[55] Jassin Kessing, Tim Tutenel, and Rafael Bidarra. 2009. Services in game worlds: A semantic approach to improve object interaction. In *International Conference on Entertainment Computing*, pages 276–281. Springer.

[56] Hakpyeong Kim and Taehoon Hong. 2024. Enhancing emotion recognition using multimodal fusion of physiological, environmental, personal data. *Expert Systems with Applications*,

249:123723.

[57] Hye-Geum Kim, Eun-Jin Cheon, Dai-Seg Bai, Young Hwan Lee, and Bon-Hoon Koo. 2018. Stress and heart rate variability: a meta-analysis and review of the literature. *Psychiatry investigation*, 15(3):235.

[58] Zheyuan Kuang, Tinghui Li, Weiwei Jiang, Sven Mayer, Flora D. Salim, Benjamin Tag, Anusha Withana, and Zhanna Sarsenbayeva. 2026. Understanding the effects of interaction on emotional experiences in vr. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

[59] Wenna Lai, Haoran Xie, Guandong Xu, Qing Li, and S Joe Qin. 2025. When llms team up: The emergence of collaborative affective computing. *arXiv preprint arXiv:2506.01698*.

[60] Peter J Lang. 1995. The emotion probe: Studies of motivation and attention. *American psychologist*, 50(5):372.

[61] Joseph J LaViola Jr, Ernst Kruijff, Ryan P McMahan, Doug Bowman, and Ivan P Poupyrev. 2017. *3D user interfaces: theory and practice*. Addison-Wesley Professional.

[62] Raymond Lavoie, Kelley Main, Corey King, and Danielle King. 2021. Virtual experience, real consequences: the potential negative emotional consequences of virtual reality gameplay. *Virtual Reality*, 25(1):69–81.

[63] Richard S Lazarus. 1991. *Emotion and adaptation*. Oxford University Press.

[64] Jeroen S Lemmens, Monika Simon, and Sindy R Sumter. 2022. Fear and loathing in vr: the emotional and physiological effects of immersive games. *Virtual reality*, 26(1):223–234.

[65] Benjamin J Li, Jeremy N Bailenson, Adam Pines, Walter J Greenleaf, and Leanne M Williams. 2017. A public database of immersive vr videos with corresponding ratings of arousal, valence, and correlations between head movements and self report measures. *Frontiers in psychology*, 8:2116.

[66] Ming Li, Junjun Pan, Yang Gao, Yang Shen, Fang Luo, Ju Dai, Aimin Hao, and Hong Qin. 2022. Neurophysiological and subjective analysis of vr emotion induction paradigm. *IEEE Transactions on Visualization and Computer Graphics*, 28(11):3832–3842.

[67] Ming Li, Junjun Pan, Yu Li, Yang Gao, Hong Qin, and Yang Shen. 2024. Multimodal physiological analysis of impact of emotion on cognitive control in vr. *IEEE Transactions on Visualization and Computer Graphics*, 30(5):2044–2054.

[68] Tinghui Li, Eduardo Velloso, Anusha Withana, and Zhanna Sarsenbayeva. 2025. Estimating the effects of encumbrance and walking on mixed reality interaction. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. Association for Computing Machinery, New York, NY, USA.

[69] Tinghui Li, Eduardo Velloso, Anusha Withana, and Zhanna Sarsenbayeva. 2025. Weight-induced consumed endurance (wice): A model to quantify shoulder fatigue with weighted objects. In *The 38th Annual ACM Symposium on User Interface Software and Technology*, UIST '25. Association for Computing Machinery, New York, NY, USA.

[70] Zechen Li, Shohreh Deldari, Linyao Chen, Hao Xue, and Flora D Salim. 2025. Sensorllm: Aligning large language models with motion sensors for human activity recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 354–379.

[71] Hailun Lian, Cheng Lu, Sunan Li, Yan Zhao, Chuangao Tang, and Yuan Zong. 2023. A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy*, 25(10):1440.

[72] Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, et al. 2025. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. *arXiv preprint arXiv:2501.16566*.

[73] Zheng Lian, Rui Liu, Kele Xu, Bin Liu, Xuefei Liu, Yazhou Zhang, Xin Liu, Yong Li, Zebang Cheng, Haolin Zuo, et al. 2025. Mer 2025: When affective computing meets large language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 13837–13842.

[74] Zheng Lian, Haiyang Sun, Licai Sun, Haoyu Chen, Lan Chen, Hao Gu, Zhuofan Wen, Shun Chen, Siyuan Zhang, Hailiang Yao, et al. 2024. Ov-mer: Towards open-vocabulary multimodal emotion recognition. *arXiv preprint arXiv:2410.01495*.

[75] Zheng Lian, Licai Sun, Yong Ren, Hao Gu, Haiyang Sun, Lan Chen, Bin Liu, and Jianhua Tao. 2026. Merbench: A unified evaluation benchmark for multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[76] Huiyi Liang, Yi Li, and Benjamin Tag. 2025. Mindfulreframer: A virtual reality system for managing exam anxiety through cognitive reappraisal. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25. Association for Computing Machinery, New York, NY, USA.

[77] Benjamin GP Linares-Vargas and Segundo E Cieza-Mostacero. 2024. Interactive virtual reality environments and emotions: A systematic review. *Virtual Reality*, 29(1):3.

[78] Stefan Liszio and Maic Masuch. 2019. Interactive immersive virtual environments cause relaxation and enhance resistance to acute stress. *Annu Rev Cyberther Telemed*, 17:65–71.

[79] Danielle Lottridge, Mark Chignell, and Aleksandra Jovicic. 2011. Affective interaction: understanding, evaluating, and designing for human emotion. *Reviews of Human Factors and Ergonomics*, 7(1):197–217.

[80] Haifeng Lu, Jiuyi Chen, Feng Liang, Mingkui Tan, Runhao Zeng, and Xiping Hu. 2025. Understanding emotional body expressions via large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1447–1455.

[81] Tiffany Luong and Christian Holz. 2022. Characterizing physiological responses to fear, frustration, and insight in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 28(11):3917–3927.

[82] Ying Ma, Qiushi Zhou, Benjamin Tag, Zhanna Sarsenbayeva, Jarrod Knibbe, and Jorge Goncalves. 2023. "hello, fellow villager!": Perceptions and impact of displaying users' locations on weibo. In *IFIP Conference on Human-Computer Interaction*, pages 511–532. Springer.

[83] Dominique Makowski, Tam Pham, Zen J Lau, Jan C Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and SH Annabel Chen. 2021. Neurokit2: A python toolbox for neurophysiological signal processing. *Behavior research methods*, 53(4):1689–1696.

[84] Guido Makransky and Lau Lilleholt. 2018. A structural equation modeling investigation of the emotional value of immersive virtual reality in education. *Educational Technology Research and Development*, 66(5):1141–1164.

[85] Jessica L Maples-Keller, Brian E Bunnell, Sae-Jin Kim, and Barbara O Rothbaum. 2017. The use of virtual reality technology in the treatment of anxiety and other psychiatric disorders. *Harvard review of psychiatry*, 25(3):103–113.

[86] Javier Marín-Morales, Carmen Llinares, Jaime Guixeres, and Mariano Alcañiz. 2020. Emotion recognition in immersive virtual reality: From statistics to affective computing. *Sensors*, 20(18):5163.

[87] Iris B Mauss and Michael D Robinson. 2009. Measures of emotion: A review. *Cognition and emotion*, 23(2):209–237.

[88] Daniel J McDuff, Javier Hernandez, Sarah Gontarek, and Rosalind W Picard. 2016. Cogcam: Contact-free measurement of cognitive stress during computer tasks with a digital camera. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4000–4004.

[89] Albert Mehrabian. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14:261–292.

[90] Albert Mehrabian and James A Russell. 1974. *An approach to environmental psychology.* the MIT Press.

[91] Axel Gedeon Mengara Mengara and Yeon-kug Moon. 2025. Cag-moe: Multimodal emotion recognition with cross-attention gated mixture of experts. *Mathematics (2227-7390)*, 13(12).

[92] Ben Meuleman and David Rudrauf. 2018. Induction and profiling of strong multi-componential emotions in virtual reality. *IEEE Transactions on Affective Computing*, 12(1):189–202.

[93] Aske Mottelson, Gustav Bøg Petersen, Klemen Lilija, and Guido Makransky. 2021. Conducting unsupervised virtual reality user studies online. *Frontiers in Virtual Reality*, 2:681482.

[94] Minxue Niu, Yara El-Tawil, Amrit Romana, and Emily Mower Provost. 2025. Rethinking emotion annotations in the era of large language models. *IEEE Transactions on Affective Computing*.

[95] James Andrew Oxley, Kristof Santa, Georg Meyer, and Carri Westgarth. 2022. A systematic scoping review of human-dog interactions in virtual and augmented reality: The use of virtual dog models and immersive equipment. *Frontiers in Virtual Reality*, 3:782023.

[96] Fotini Patrona, Anargyros Chatzitofis, Dimitrios Zarpalas, and Petros Daras. 2018. Motion analysis: Action detection, recognition and evaluation based on motion capture data. *Pattern Recognition*, 76:612–622.

[97] Katarina Pavic, Laurence Chaby, Thierry Gricourt, and Dorine Vergilino-Perez. 2023. Feeling virtually present makes me happier: The influence of immersion, sense of presence, and video contents on positive emotion induction. *Cyberpsychology, Behavior, and Social Networking*,

26(4):238–245.

[98] Joann Peck and Suzanne B Shu. 2009. The effect of mere touch on perceived ownership. *Journal of consumer Research*, 36(3):434–447.

[99] John P Pollak, Phil Adams, and Geri Gay. 2011. Pam: a photographic affect meter for frequent, in situ measurement of affect. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 725–734.

[100] Dominic Potts, Zoe Broad, Tarini Sehgal, Joseph Hartley, Eamonn O'Neill, Crescent Jicol, Christopher Clarke, and Christof Lutteroth. 2024. Sweating the details: Emotion recognition and the influence of physical exertion in virtual reality exergaming. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

[101] Dominic Potts, Miloni Gada, Aastha Gupta, Kavya Goel, Klaus Philipp Krzok, Genevieve Pate, Joseph Hartley, Mark Weston-Arnold, Jakob Aylott, Christopher Clarke, et al. 2025. Retrosketch: A retrospective method for measuring emotions and presence in virtual reality. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–25.

[102] Giuseppe Riva, Fabrizia Mantovani, Claret Samantha Capideville, Alessandra Preziosa, Francesca Morganti, Daniela Villani, Andrea Gaggioli, Cristina Botella, and Mariano Alcañiz. 2007. Affective interactions using virtual reality: the link between presence and emotions. *Cyberpsychology & behavior*, 10(1):45–56.

[103] Giuseppe Riva, Brenda K Wiederhold, and Fabrizia Mantovani. 2019. Neuroscience of virtual reality: from virtual exposure to embodied medicine. *Cyberpsychology, behavior, and social networking*, 22(1):82–96.

[104] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

[105] Zhanna Sarsenbayeva, Gabriele Marini, Niels van Berkel, Chu Luo, Weiwei Jiang, Kangning Yang, Greg Wadley, Tilman Dingler, Vassilis Kostakos, and Jorge Goncalves. 2020. Does smartphone use drive our emotions or vice versa? a causal analysis. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–15.

[106] Zhanna Sarsenbayeva, Niels van Berkel, Danula Hettiachchi, Weiwei Jiang, Tilman Dingler, Eduardo Velloso, Vassilis Kostakos, and Jorge Goncalves. 2019. Measuring the effects of stress on mobile interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(1):1–18.

[107] Anthony Scavarelli, Ali Arya, and Robert J Teather. 2021. Virtual reality and augmented reality in social learning spaces: a literature review. *Virtual reality*, 25(1):257–277.

[108] Benjamin Schöne, Joanna Kisker, Rebecca Sophia Sylvester, Elise Leila Radtke, and Thomas Gruber. 2023. Library for universal virtual reality experiments (luvre): A standardized immersive 3d/360 picture and video database for vr based research. *Current Psychology*, 42(7):5366–5384.

[109] Yuntao Shou, Tao Meng, Wei Ai, and Keqin Li. 2025. Multimodal large language models meet multimodal emotion recognition and reasoning: A survey. *arXiv preprint arXiv:2509.24322*.

[110] Ewa Siedlecka and Thomas F Denson. 2019. Experimental methods for inducing basic emotions: A qualitative review. *Emotion Review*, 11(1):87–97.

[111] Richard Skarbez, Frederick P Brooks, and Mary C Whitton. 2020. Immersion and coherence: Research agenda and early results. *IEEE transactions on visualization and computer graphics*, 27(10):3839–3850.

[112] Mel Slater and Maria V Sanchez-Vives. 2016. Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI*, 3:236866.

[113] Rukshani Somarathna, Tomasz Bednarz, and Gelareh Mohammadi. 2022. Virtual reality for emotion elicitation–a review. *IEEE Transactions on Affective Computing*, 14(4):2626–2645.

[114] Liana Spytska. 2024. The use of virtual reality in the treatment of mental disorders such as phobias and post-traumatic stress disorder. *SSM-Mental Health*, 6:100351.

[115] Nazmi Sofian Suhaimi, James Mountstephens, and Jason Teo. 2020. Eeg-based emotion recognition: a state-of-the-art review of current trends and opportunities. *Computational intelligence and neuroscience*, 2020(1):8875426.

[116] Benjamin Tag, Zhanna Sarsenbayeva, Anna L Cox, Greg Wadley, Jorge Goncalves, and Vassilis Kostakos. 2022. Emotion trajectories in smartphone use: Towards recognizing emotion regulation in-the-wild. *International Journal of Human-Computer Studies*, 166:102872.

[117] Jan-Philipp Tauscher, Fabian Wolf Schottky, Steve Grogorick, Paul Maximilian Bittner, Maryam Mustafa, and Marcus Magnor. 2019. Immersive eeg: evaluating electroencephalography in virtual reality. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 1794–1800. IEEE.

[118] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 6558–6569.

[119] Deltcho Valtchanov and Colin Ellard. 2010. Physiological and affective responses to immersion in virtual reality: effects of nature and urban settings. *Journal of CyberTherapy & Rehabilitation (JCR)*, 3(4).

[120] Nadine Wagener, Arne Kiesewetter, Leon Reicherts, Paweł W Woźniak, Johannes Schöning, Yvonne Rogers, and Jasmin Niess. 2024. Moodshaper: A virtual reality experience to support managing negative emotions. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, pages 2286–2304.

[121] Wenbin Wang, Liang Ding, Li Shen, Yong Luo, Han Hu, and Dacheng Tao. 2024. Wisdom: Improving multimodal sentiment analysis by fusing contextual world knowledge. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 2282–2291.

[122] Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

[123] Claudia AF Wascher. 2021. Heart rate as a measure of emotional arousal in evolutionary biology. *Philosophical Transactions of the Royal Society B*, 376(1831):20200479.

[124] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063.

[125] David Watson, David Wiese, Jatin Vaidya, and Auke Tellegen. 1999. The two general activation systems of affect: Structural findings, evolutionary considerations, and psychobiological evidence. *Journal of personality and social psychology*, 76(5):820.

[126] Qinglan Wei, Ruiqi Xue, Hongjiang Xiao, Yuan Zhang, Long Ye, and Yutian Wang. 2025. Mimicking the mavens: agent-based opinion synthesis and emotion prediction for social media influencers. *Journal of Social Computing*, 6(3):221–238.

[127] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 143–146.

[128] Jialan Xie, Yutong Luo, Shiyuan Wang, and Guangyuan Liu. 2024. Electroencephalography-based recognition of six basic emotions in virtual reality environments. *Biomedical Signal Processing and Control*, 93:106189.

[129] Jun Xie, Yingjian Zhu, Feng Chen, Zhenghao Zhang, Xiaohui Fan, Hongzhu Yi, Xinming Wang, Chen Yu, Yue Bi, Zhaoran Zhao, et al. 2025. More is better: A moe-based emotion recognition framework with human preference alignment. In *Proceedings of the 3rd International Workshop on Multimodal and Responsible Affective Computing*, pages 2–7.

[130] Tianhua Xie, Mingliang Cao, and Zhigeng Pan. 2020. Applying self-assessment manikin (sam) to evaluate the affective arousal effects of vr games. In *Proceedings of the 2020 3rd International Conference on Image and Graphics Processing*, pages 134–138.

[131] Dingkang Yang, Zhaoyu Chen, Yuzheng Wang, Shunli Wang, Mingcheng Li, Siao Liu, Xiao Zhao, Shuai Huang, Zhiyan Dong, Peng Zhai, et al. 2023. Context de-confounded emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19005–19015.

[132] Kangning Yang, Benjamin Tag, Yue Gu, Chaofan Wang, Tilman Dingler, Greg Wadley, and Jorge Goncalves. 2022. Mobile emotion recognition via multiple physiological signals using convolution-augmented transformer. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pages 562–570.

[133] Kangning Yang, Chaofan Wang, Yue Gu, Zhanna Sarsenbayeva, Benjamin Tag, Tilman Dingler, Greg Wadley, and Jorge Goncalves. 2021. Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition. *IEEE Transactions on Affective Computing*, 14(2):1082–1097.

[134] Kangning Yang, Chaofan Wang, Zhanna Sarsenbayeva, Benjamin Tag, Tilman Dingler, Greg Wadley, and Jorge Goncalves. 2021. Benchmarking commercial emotion detection systems using realistic distortions of facial image datasets. *The visual computer*, 37(6):1447–1466.

[135] Difeng Yu, Tilman Dingler, Eduardo Velloso, and Jorge Goncalves. 2024. Object selection and manipulation in vr headsets: Research challenges, solutions, and success measurements. *ACM Computing Surveys*, 57(4):1–34.

[136] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10790–10797.

[137] Xiaofan Yu, Lanxiang Hu, Benjamin Reichman, Dylan Chu, Rushil Chandrupatla, Xiyuan Zhang, Larry Heck, and Tajana S Rosing. 2025. Sensorchat: Answering qualitative and quantitative questions during long-term multimodal sensor interactions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3):1–35.

[138] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1103–1114.

[139] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.

[140] Chengwen Zhang, Yaohui Liu, and Bo Cheng. 2025. A moe multimodal graph attention network framework for multimodal emotion recognition. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

[141] Emma Yann Zhang, Zhigeng Pan, and Adrian David Cheok. 2025. Emotion recognition using affective touch: A survey. *IEEE Transactions on Affective Computing*.

[142] Haiwei Zhang, Jiqing Zhang, Bo Dong, Pieter Peers, Wenwei Wu, Xiaopeng Wei, Felix Heide, and Xin Yang. 2023. In the blink of an eye: Event-based emotion recognition. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11.

[143] Haochen Zhang, Yuyang Dong, Chuan Xiao, and Masafumi Oyamada. 2023. Large language models as data preprocessors. *arXiv:2308.16361*.

[144] Yiqun Zhang, Xiaocui Yang, Xingle Xu, Zeran Gao, Yijie Huang, Shiyi Mu, Shi Feng, Daling Wang, Yifei Zhang, Kaisong Song, et al. 2024. Affective computing in the era of large language models: A survey from the nlp perspective. *arXiv preprint arXiv:2408.04638*.

[145] Yangyang Zhou, Xin Kang, and Fuji Ren. 2023. Prompt consistency for multi-label textual emotion detection. *IEEE Transactions on Affective Computing*, 15(1):121–129.

[146] Yitong Zhu, Lei Han, GuanXuan Jiang, PengYuan Zhou, and Yuyang Wang. 2025. Hierarchical moe: Continuous multimodal emotion recognition with incomplete and asynchronous inputs. *arXiv preprint arXiv:2508.02133*.

APPENDIX A

Appendix A

A.1 Supplementary Tables

TABLE A.1. Demographic details and VR experience of participants in the Interactive and Non-Interactive conditions.

INTERACTION	**Gender**		**Age**		**VR Exp.**			
	M	**F**	**M**	**SD**	**Never**	**Daily**	**Weekly**	**Monthly**
Non-Interactive	21	21	23.79	3.21	13	7	10	12
Interactive	21	21	24.31	2.87	11	7	11	13

TABLE A.2. Mean and Median SAM ratings for the *Surrounded by Elephants* VR scene.

Condition	**Valence**		**Arousal**		**Dominance**	
	M	Med	M	Med	M	Med
Previous Study [65]	5.94	NR	5.56	NR	NR	NR
360° Video	5.42	5.50	5.67	6.00	5.08	4.50
VR Scene	6.71	7.00	5.54	6.00	6.29	7.00

NR = Not reported in the previous study.

TABLE A.3. Fixed-effects estimates from linear mixed-effects models for high-frequency HRV (HF) and heart rate (HR). Each outcome was modelled as *Outcome* $\sim$ Condition * Scene + (1|Participant). Values are reported as Estimate (SE) with 95% CI.

Parameter	**HF**			**HR**		
	Estimate	SE	95% CI	Estimate	SE	95% CI
Intercept	**9.29*****	0.12	[9.06, 9.52]	**77.67*****	1.99	[73.74, 81.59]
Interactive	**0.69*****	0.17	[0.36, 1.02]	**-9.25****	2.82	[-14.80, -3.70]
Puppies	**0.31***	0.12	[0.06, 0.55]	-1.55	1.93	[-5.36, 2.25]
Jetty at Lake	0.13	0.13	[-0.12, 0.38]	3.55	1.99	[-0.37, 7.47]
Solitary Confinement	**0.44*****	0.13	[0.19, 0.69]	-0.32	1.96	[-4.18, 3.54]
Shouting Man with Gun	**0.30***	0.13	[0.05, 0.55]	0.09	1.98	[-3.79, 3.98]
Surrounded by Elephants	-0.01	0.13	[-0.26, 0.23]	1.74	1.96	[-2.11, 5.59]
Interactive $\times$ Puppies	-0.22	0.18	[-0.56, 0.13]	1.46	2.75	[-3.94, 6.86]
Interactive $\times$ Jetty at Lake	-0.04	0.18	[-0.39, 0.31]	-2.74	2.79	[-8.22, 2.74]
Interactive $\times$ Solitary Confinement	-0.34	0.18	[-0.69, 0.01]	-0.39	2.78	[-5.84, 5.07]
Interactive $\times$ Shouting Man with Gun	**-0.49****	0.18	[-0.84, -0.14]	3.87	2.79	[-1.60, 9.35]
Interactive $\times$ Surrounded by Elephants	-0.24	0.18	[-0.58, 0.11]	2.71	2.75	[-2.70, 8.12]

Notes. Stars denote significance: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

TABLE A.4. Fixed-effects estimates from linear mixed-effects models for SCR Count and SCR amplitude. Each outcome was modelled as *Outcome* $\sim$ Condition * Scene + (1|Participant). Values are reported as Estimate (SE) with 95% CI.

Parameter	**SCR Count**			**SCR Amplitude**		
	Estimate	SE	95% CI	Estimate	SE	95% CI
Intercept	**16.67*****	2.69	[11.35, 21.98]	**347.31***	136.68	[77.05, 617.57]
Interactive	5.50	3.81	[-2.01, 13.01]	188.73	193.29	[-193.48, 570.93]
Puppies	0.67	2.58	[-4.40, 5.74]	113.03	102.73	[-88.92, 314.98]
Jetty at Lake	0.98	2.58	[-4.10, 6.05]	4.39	102.73	[-197.56, 206.34]
Solitary Confinement	-2.60	2.58	[-7.67, 2.48]	184.08	102.73	[-17.87, 386.03]
Shouting Man with Gun	0.28	2.64	[-4.90, 5.47]	187.77	105.04	[-18.73, 394.27]
Surrounded by Elephants	0.50	2.58	[-4.57, 5.57]	-3.87	102.73	[-205.82, 198.08]
Interactive $\times$ Puppies	0.77	3.66	[-6.43, 7.96]	1.02	145.82	[-285.65, 287.68]
Interactive $\times$ Jetty at Lake	-4.74	3.65	[-11.91, 2.43]	233.83	145.28	[-51.77, 519.43]
Interactive $\times$ Solitary Confinement	6.37	3.66	[-0.83, 13.57]	-141.12	145.82	[-427.78, 145.55]
Interactive $\times$ Shouting Man with Gun	-1.57	3.72	[-8.88, 5.73]	-47.73	147.99	[-338.65, 243.19]
Interactive $\times$ Surrounded by Elephants	4.79	3.65	[-2.39, 11.96]	61.25	145.28	[-224.35, 346.85]

Notes. Stars denote significance: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

A.2 Questionnaire and Interview

A.2.1 Questionnaire

- What is your gender identity?

  ☐ Woman ☐ Man ☐ Non-binary/gender diverse ☐ My gender identity is not listed
- What is your current age (in years)?
- How often do you use VR headsets?

  ☐ Never ☐ Daily ☐ Weekly ☐ Monthly

A.2.2 Interview

- How would you rate your overall experience with the VR scenes?

  (Not enjoyable at all) 0 1 2 3 4 5 (Neutral) 6 7 8 9 10 (Extremely enjoyable)
- Did you experience any discomfort or dizziness during or after the VR experience?
- Which VR scene did you find the most emotionally impactful?

  ☐ Tunnel ☐ Puppies ☐ Jetty at Lake

  ☐ Solitary Confinement ☐ Shouting Man with Gun ☐ Surrounded by Elephants
- Why do you think that scene is the most emotionally impactful?
- Which VR scene did you find the least emotionally impactful?

  ☐ Tunnel ☐ Puppies ☐ Jetty at Lake

  ☐ Solitary Confinement ☐ Shouting Man with Gun ☐ Surrounded by Elephants
- Why do you think that scene is not as emotionally impactful as other scenes?

A.2.3 Self-Assessment Manikin (SAM)

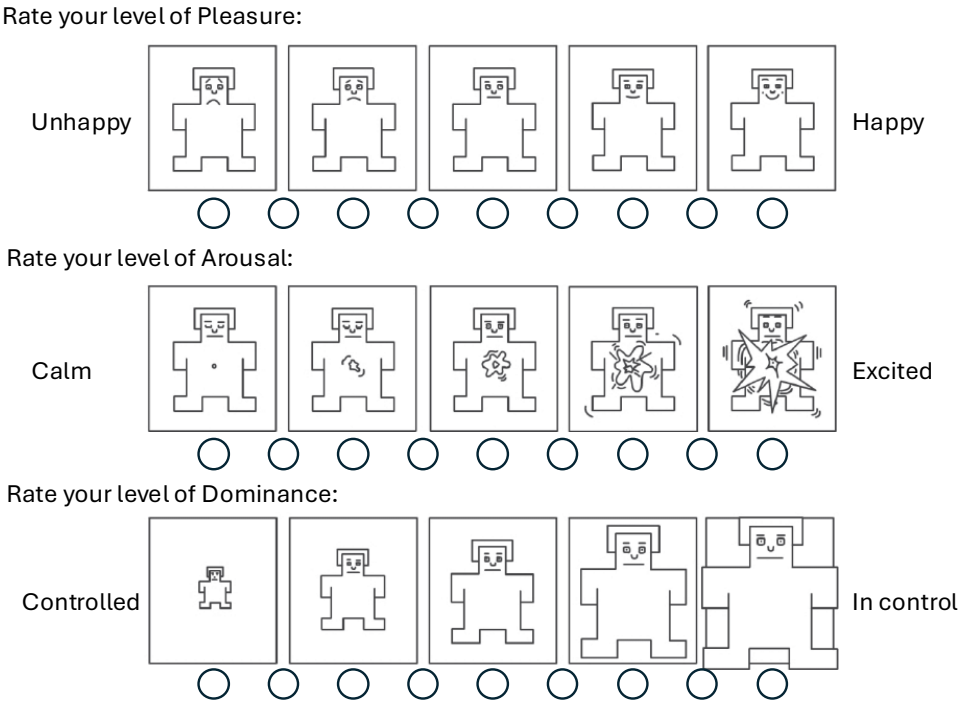


FIGURE A.1. The self-assessment manikin (SAM) questionnaire.

A.3 Interactive Objects

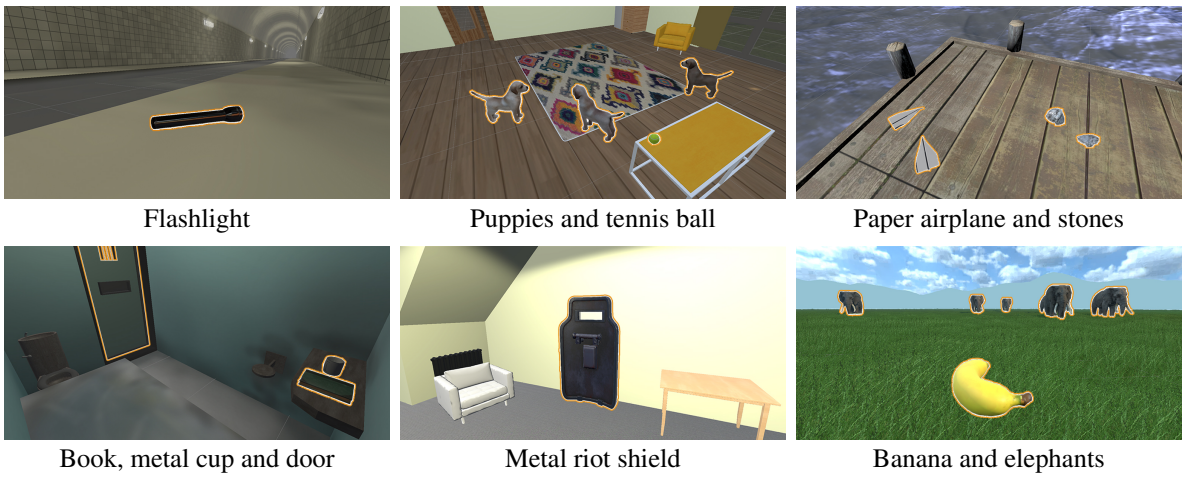


FIGURE A.2. Interactive objects with yellow outlines in the six VR scenes.