Through the Eyes of Emotion
A Multi-faceted Eye Tracking Dataset for Emotion Recognition in Virtual Reality

Yang, Tongyun; Regmi, Bishwas; Du, Lingyu; Bulling, Andreas; Zhang, Xucong; Lan, Guohao

DOI
10.1145/3749545

Publication date
2025

Document Version
Final published version

Published in
Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies

Citation (APA)
Yang, T., Regmi, B., Du, L., Bulling, A., Zhang, X., & Lan, G. (2025). Through the Eyes of Emotion: A Multi-faceted Eye Tracking Dataset for Emotion Recognition in Virtual Reality. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *9*(3), Article 143. https://doi.org/10.1145/3749545

Through the Eyes of Emotion: A Multi-faceted Eye Tracking Dataset for Emotion Recognition in Virtual Reality

TONGYUN YANG[†], Delft University of Technology, The Netherlands
BISHWAS REGMI[†], Delft University of Technology, The Netherlands
LINGYU DU, Delft University of Technology, The Netherlands
ANDREAS BULLING, University of Stuttgart, Germany
XUCONG ZHANG, Delft University of Technology, The Netherlands
GUOHAO LAN, Delft University of Technology, The Netherlands

Virtual Reality (VR) is transforming cognitive and psychological research by enabling immersive simulations that elicit authentic emotional responses. The high demand for VR-based emotion recognition is also evident in fields such as mental healthcare, education, and entertainment, where understanding users' emotional states can enhance user experience and system effectiveness. However, the lack of comprehensive datasets hinders progress in VR-based emotion recognition. In this paper, we present a comprehensive, multi-faceted eye-tracking dataset collected from 26 participants using 28 emotional video stimuli rendered in a custom virtual environment. Our dataset is the first to incorporate high-frame-rate periocular videos, capturing subtle motions, such as micro-expressions and eyebrow shifts, which are critical for emotion analysis. Additionally, it includes high-frequency eye-tracking data, offering gaze direction and pupil dynamics at four times the frequency of existing datasets. Our dataset is also unique in providing emotion annotations according to Ekman's emotion model and, as such, offering experiments impossible using existing datasets. Our benchmark evaluations show that fusing the multi-faceted eye-tracking signals in our dataset significantly improves emotion recognition accuracy. As such, our work has the potential to significantly accelerate and enable entirely new research on emotion-aware VR applications.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing design and evaluation methods**; **Virtual reality**.

Additional Key Words and Phrases: Emotion Recognition, Virtual Reality, Eye Tracking, Dataset

ACM Reference Format:
Tongyun Yang, Bishwas Regmi, Lingyu Du, Andreas Bulling, Xucong Zhang, and Guohao Lan. 2025. Through the Eyes of Emotion: A Multi-faceted Eye Tracking Dataset for Emotion Recognition in Virtual Reality. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 3, Article 143 (September 2025), 41 pages. https://doi.org/10.1145/3749545

[†] Co-author equal contribution.
The dataset and software tool introduced in this paper are available via: https://github.com/MultiRepEyeVR/Through-the-Eyes-of-Emotion

Authors' Contact Information: Tongyun Yang, Delft University of Technology, Delft, The Netherlands, tonguyunyang@tudelft.nl; Bishwas Regmi, Delft University of Technology, Delft, The Netherlands, bishwas182@gmail.com; Lingyu Du, Delft University of Technology, Delft, The Netherlands, Lingyu.Du@tudelft.nl; Andreas Bulling, University of Stuttgart, Stuttgart, Germany, andreas.bulling@vis.uni-stuttgart.de; Xucong Zhang, Delft University of Technology, Delft, The Netherlands, xucong.zhang@tudelft.nl; Guohao Lan, Delft University of Technology, Delft, The Netherlands, g.lan@tudelft.nl.

ACM 2474-9567/2025/9-ART143
https://doi.org/10.1145/3749545

1 INTRODUCTION

In recent years, Virtual Reality (VR) has received substantial attention from academia and industry. In VR, users can see, hear, and interact with virtual 3D environments, experiencing a sense of presence that isolates them from the real world and fosters deep physical and emotional immersion [77, 111]. Due to its immersive and interactive nature, VR has become a powerful tool for cognitive and psychological research [64, 68]. By simulating environments and scenarios that elicit genuine emotional and cognitive responses, VR enables researchers to observe and analyze emotional cues in realistic yet controlled environments [24, 51, 64].

Studies have shown that VR environments can effectively trigger and assess a broad spectrum of emotions [63, 84, 107], leading to applications across fields, such as mental healthcare [31, 40], education [2, 14], and entertainment [65]. For instance, in mental healthcare, VR was used in exposure therapy to treat phobias and anxiety disorders [23, 26, 31]. Similarly, educational institutions employed VR to provide immersive learning experiences, enhancing knowledge retention and comprehension [96, 110]. In the entertainment industry, VR has created more emotionally engaging storytelling experiences [19, 89].

In addition to its benefits for effective emotion elicitation, awareness of the VR user's emotional states and responses to VR content is crucial for various applications in affective computing [85]. For instance, in VR-based social interaction [37, 92], the emotional state of the user can assist the creation of expressive avatars that dynamically reflect their emotional expressions [3, 7, 82, 118, 128]. Such emotionally responsive avatars can enhance social presence and empathy, making interactions more authentic and engaging. Recognizing patients' emotional responses is also critical in the healthcare domain, such as VR-based mental health treatments [31]. For instance, the real-time emotional feedback from the patients can be used to tailor therapeutic experiences to individual needs and to assess progress in reducing symptoms of anxiety, depression, or post-traumatic stress disorder [31].

As with many human context-sensing and recognition applications, developing robust and highly accurate emotion recognition algorithms relies on the availability of suitable training datasets. However, collecting large-scale emotion datasets in VR with accurate annotations is notoriously challenging [5, 123], and thus, only a few such datasets are available [42, 63, 111, 122]. Most of them rely on external physiological and behavioral data, such as electrocardiogram, galvanic skin response, and heart rate, which are not integrated into most commercial VR devices, making these approaches intrusive and limiting their scalability in practical VR applications. Moreover, unlike conventional, non-immersive settings that have access to a wide variety of validated audiovisual stimuli [30, 78], standardized emotional stimuli for VR are rare [51, 63, 111]. The scarcity of validated VR stimuli adds complexity to creating large-scale emotion recognition datasets, which is essential for advancing research in this emerging area.

To address this gap, we present a comprehensive, multi-faceted eye-tracking dataset, collected from 26 subjects exposed to 28 emotional video stimuli rendered in a virtual environment customized to enhance user's immersive experience. Alongside the dataset, we introduce a purpose-built data collection tool that streamlines synchronized capture of gaze and contextual data, specifically tailored for emotion recognition research in VR settings. Our work advances the state of the art in three distinct ways:

- We present the first emotion recognition dataset in VR that incorporates periocular videos captured by binocular near-eye cameras. These cameras record subtle movement cues, such as micro-expressions, pupil constriction and dilation, and eyebrow movements, that are closely linked to affective states and emotional shifts, enabling fine-grained analysis of users' emotional and mental states.
- Our dataset includes high-resolution binocular eye-tracking data featuring 2D gaze direction, pupil diameter, and pupil positions, captured at four times the frequency of state-of-the-art datasets [111]. This enhanced detail can enable a more precise analysis of eye movement-related features, which is critical for understanding emotional and cognitive states in VR.

- Our dataset uniquely applies Ekman's basic emotion model [29] for emotion annotation in VR. Unlike existing datasets [42, 63, 111, 122] that primarily based on the Circumplex Model of Affect [93], our dataset provides a discrete emotion framework that enables real-time emotion recognition in immersive environments. This supports a range of future applications, including adaptive training and education, emotionally responsive virtual assistants, dynamic game engines, and personalized therapeutic experiences. These use cases benefit from the interpretability and responsiveness offered by categorical emotion modeling, which is more challenging to achieve with continuous valence-arousal representations.

To demonstrate the value and validity of our dataset, we present a benchmark evaluation that highlights its potential to enable highly accurate emotion recognition. Specifically, we demonstrate that by fusing the new periocular video data incorporated in our dataset with conventional eye-tracking signals, we can significantly improve the emotion recognition performance in cross-session scenarios (with unseen video stimuli) and few-shot learning settings. The dataset and software tool introduced in this paper are available via: https://github.com/MultiRepEyeVR/Through-the-Eyes-of-Emotion.

2 RELATED WORK

2.1 Emotion Models

Existing methods for measuring emotion are broadly classified into categorical and dimensional models [81]. Categorical models require selecting a single emotion from a predefined set to best represent the feeling conveyed, with examples including Ekman's six basic emotions [29] and Izard's ten core emotions [49]. By contrast, dimensional models use quantitative measures through multidimensional scaling, where each dimension reflects a specific feature of human emotion, and their combination provides an interpretation of the emotional state. For instance, the widely adopted Circumplex Model of Affect, introduced by Russell [93], uses a two-dimensional framework in which emotions are mapped within a circular space to represent valence (pleasantness) and arousal (intensity or activation level). In order to differentiate closely related emotions, the pleasure-arousal-dominance (PAD) model introduces dominance as a third dimension [76]. To quantify these dimensional scales, researchers employ tools such as the self-assessment manikin (SAM) [10], or Feeltrace [20].

In this study, we adopt Ekman's model and consider seven basic emotions, i.e., *Happiness, Sadness, Fear, Disgust, Anger, Surprise*, and *Neutral*. We also incorporate the emotional intensity rating [108, 138] to capture the strength or degree of each emotion experienced by our research participants. This intensity rating enables us to move beyond basic categorization. By capturing variations in emotion intensity, our dataset provides a foundation for future research to explore the relationship between human visual behaviors, such as gaze, pupil dynamics, and periocular micro-movements, and subtle differences in emotional intensity.

2.2 Methods for Emotion Recognition

Emotion Recognition Using a Single Modality. Early work in emotion recognition primarily focused on uni-modal approaches, utilizing a single modality, such as facial expressions, voice, or physiological signals, to train machine learning models for emotion recognition. For instance, Siqueira et al. [103] leverage facial expressions as input and achieve 87.15% accuracy on the FER+ dataset using a convolutional neural network (CNN)-based classifier. In voice-based emotion recognition, recurrent neural networks (RNNs), particularly long short-term memory (LSTM) networks, have shown strong performance in capturing temporal dependencies. For example, Zhao et al. [131] achieve over 95% accuracy across seven emotions on the Berlin EmoDB dataset [12] in both speaker-dependent and speaker-independent scenarios. Similarly, emotion recognition methods using physiological signals have benefited from deep learning advancements. Zhong et al. [134] achieve 73.84% accuracy in classifying four emotions on the SEED-IV dataset using graph neural networks (GNNs), which effectively model the spatial relations among brain regions by using the inherent graph structure of EEG electrode placements.

Emotion Recognition Using Multi-modality. While uni-modal approaches have made significant progress in emotion recognition, there is a growing trend toward multi-modal methods, which leverage complementary modalities for robust results. For instance, Tan et al. [112] demonstrate an 83.33% recognition rate by integrating facial expressions with EEG signals. Wampfer et al. [116] combine touch events during smartphone keystrokes with inertial sensor data to generate two-dimensional heat maps, which are then processed by a CNN to predict users' affective states. In an in-vehicle setting, Bethge et al. [8] improve driver emotion prediction by incorporating contextual data from smartphones, such as road conditions, visual scenes, audio, weather, and car speed. This integration yields a 7% accuracy improvement over facial expression only methods. In the setting of passive social media usage, Gebhardt et al.[36] explore emotion recognition by combining behavioral data (phone interaction, motion, touch, facial expressions, and eye tracking) with physiological signals (ECG, PPG, and EDA). Their classifier, trained with multi-modal data, accurately detects up to eight emotional states, achieving a peak accuracy of 83%. More recently, Park et al. [83] leverage multi-modal data, including audio, ECG, EEG, EDA, acceleration, and temperature signals from wearable devices, to assess workers' emotional workload in emotional labor settings. By using statistical features and standard machine learning methods, they achieve up to 87% accuracy for both binary and three-class emotion state classifications.

2.3 Datasets for Emotion Recognition

As with many context-sensing and recognition applications, the development of robust and highly accurate emotion recognition systems depends heavily on the availability of comprehensive datasets. However, collecting large-scale emotion data with precise annotations is notoriously challenging [5, 123]. There are a few datasets in the literature, collected either in conventional, non-immersive interaction setups [52, 55, 78, 106], or immersive environments [42, 111, 123, 133] (as considered in our work). We review these works briefly below.

Datasets Collected with Non-immersive Setup. There are many datasets collect physiological and behavioral data using dedicated sensors, such as eye trackers, galvanic skin response (GSR), electrocardiograms (ECG), and electroencephalograms (EEG), while subjects engage with affective stimuli [52, 55, 78, 106]. For instance, the MAHNOB-HCI dataset [106] is a multi-modal emotion recognition dataset that collects facial expressions, audio signals, eye-gaze data, and physiological signals (ECG and skin conductivity), and respiration from 27 participants when they watch 20 emotionally charged videos. Similarly, the DREAMER dataset [52] contains EEG and ECG signals collected during emotion elicitation via audio-visual stimuli, with data from 23 participants and self-assessments of valence, arousal, and dominance after each stimulus.

There are also datasets that focus on facial expressions for emotion recognition without relying on wearable sensors. FER+ [6], for example, includes 28,709 internet-sourced images annotated with multi-label classifications across seven emotions. Similarly, AffectNet [78] features 440,000 images manually annotated with single-label emotions across the same seven categories as FER+. Unlike static images, video provides a temporal dimension for emotion recognition, enabling the capture of dynamic expressions. For instance, the MAFW dataset [66] contains multi-modal clips from sources such as movies, TV dramas, and short videos, where subjects express a range of emotions in diverse contexts, adding valuable contextual information for more nuanced emotion recognition.

Datasets Collected in VR. Despite the substantial body of work on studying user emotion and affect in VR [68, 71, 125], to the best of our knowledge, **there are very limited publicly VR datasets exist for emotion recognition** [42, 63, 111, 122]. These works often incorporate external sensors to capture behavioral and physiological data, such as heart rate variability (HRV), electrodermal activity (EDA), gaze patterns, and brain wave signals [42, 111, 122], when the participants are presented with stimuli designed to evoke specific emotions in an immersive environment. For example, the VREED dataset [111] includes eye movement data (gaze estimation and pupil diameter), ECG, galvanic skin response, and self-reported emotional states from 34 participants while they are viewing 360° videos in a VR headset. Similarly, the PEM360 dataset [42] captures similar signals to VREED, with the addition of head movements and heart rate, from 32 participants watching eight videos. The CEAP-360VR dataset [122] is collected from 32 participants and includes self-reported levels of motion sickness and presence, in addition to eye movement and ECG signals, offering insights into user comfort in immersive environments such as VR.

2.4 Discussion: Rationale for a New Dataset

Compared to existing VR emotion recognition datasets, our work stands out in four key ways.

First, to the best of our knowledge, our dataset is **the first one to incorporate periocular videos captured by two near-eye cameras**. These videos record subtle motions, such as micro-expressions and eyebrow shifts, that are valuable for emotion recognition and cognitive studies. The effectiveness of the periocular videos has been shown in previous works with the head-mounted eye tracker [129, 132], yet it has not been studied in the immersive VR settings considered here. Our experiment shows that utilizing the periocular videos can significantly improve emotion recognition performance.

Second, **our dataset provides high-resolution, multi-faceted eye-tracking data**, including 2D gaze direction, pupil diameter, and pupil position at four times the frequency of current state-of-the-art datasets [111, 122]. It is distinct from existing datasets that include only pre-processed eye-tracking data [111, 122], such as gaze and fixation patterns, for emotion recognition in VR [43, 111, 122]. Our dataset enables future studies to explore the rich information within the eye movement information.

Third, **ours is the first to apply Ekman's basic model for emotion recognition in VR**. Although all existing datasets [42, 63, 111, 122] adopt the Circumplex Model of Affect [93] as their emotion model, there is ongoing debate about whether emotional expressions are universally recognized or influenced by cultural context [28, 73, 94]. Nonetheless, we believe our dataset offers a valuable alternative for studying emotion recognition in VR. By leveraging Ekman's core emotion types, algorithms developed using our dataset can directly map recognized emotions to real-world applications, making it highly suitable for cognitive-aware applications in VR. For instance, detected core emotional states of VR users can be used to help generate "expressive avatars" [7] that exhibit realistic emotional expressions and responses [3, 82, 118, 128]. By contrast, achieving this level of emotion expressive accuracy is far more challenging when relying on the complex, continuous dimensions of valence and arousal in the Circumplex model.

Fourth, **we provide an open-source dataset collection software tool in VR headset** using only the standard onboard sensors. The developed data collection pipeline can collect eye movement data as well as periocular video data simultaneously, which can be used for data collection of variant tasks such as cognitive load estimation, attention estimation, human-AI interaction, etc. Distinct from all existing datasets that rely on additional external sensors [42, 111, 122], our software tool enhances compatibility with a broad range of commercial extended reality (XR) hardware, improving the practicality and scalability of dataset collection.

3 DESIGN OF THE DATA COLLECTION SYSTEM

In this work, we develop an end-to-end system to facilitate future multi-modal data collection and annotation for VR-based emotion and affective computing studies.

3.1 Hardware

The VR platform we used for data collection is the VIVE Pro EYE VR headset, which integrates eye-tracking alongside conventional sensors like IMUs and microphones. This setup enables simultaneous capture of multiple signal modalities on a single device, i.e., eye movement [111], head movement [97, 121], and voice [30, 83], that are known to be useful for emotion recognition. Moreover, the VIVE Pro EYE is a widely available, high-performance VR device that delivers professional-grade graphics and audio. Its dual OLED displays, with a combined resolution of 2880 × 1600 pixels and 615 PPI (pixels per inch), provide vivid colors and sharp contrast. The integrated Hi-Res certified headphones deliver 3D spatial sound for an immersive audio experience. Together, these features enhance the ability of our system to elicit genuine emotional responses of the participant by creating a deeply immersive virtual environment both physically and emotionally.

However, extracting raw periocular images from the integrated eye-tracking module of VIVE Pro EYE requires the Tobii XR SDK license. As shown in Figure 14 (in Appendix D), to overcome this, we integrate the VR headset with an external eye-tracking add-on from Pupil Labs, allowing us to collect raw periocular images at a high frame rate. Specifically, the Pupil Lab eye tracker includes two infrared near-eye cameras that capture images of the periocular area at 120Hz with a resolution of 400×400 pixels. The eye tracker is connected to the VR headset via the onboard USB port, which ensures stable data transmission and sufficient bandwidth for dual eye-tracking video streaming. We are particularly interested in periocular images, as studies in affective computing and psychology [13, 91] have shown that the periocular area, i.e., the region surrounding the eyes, including the eyebrows and upper cheeks, contains rich information about emotional states. For instance, changes in shape, wrinkles, and movements around the eyes can effectively signal emotional states. As demonstrated in our benchmark study (Section 5), integrating these periocular features with conventional eye-tracking signals, such as pupil dilation and gaze direction, significantly enhances emotion recognition performance.

3.2 Software

We develop a software tool for data collection and labeling, consisting of three main components: the virtual environment, the data recording module, and the data labeling module.

3.2.1 Virtual Environment. The virtual environment is where participants are immersed while wearing the VR headset during data collection. Instead of using the default video player, we custom-designed the virtual environment to create a controlled setting that enhances emotion elicitation. Specifically, the main challenge is the lack of affective 360° videos in the literature that can reliably elicit Ekman's basic emotions [29]. Existing VR emotion datasets primarily use panoramic videos focused on dimensional emotion states, such as valence and arousal [51, 63, 64, 111, 122, 123]. To incorporate conventional non-panoramic 2D videos that have proven effective in eliciting the six basic emotions, we design the virtual environment to render these videos in a way that ensures participant comfort while preserving the effectiveness of the emotional stimuli. Two key design considerations were made. First, the viewing experience of non-panoramic videos is crafted to feel as natural as possible within the virtual environment, allowing participants to become fully immersed. Second, the video stimuli are presented to enable participants to easily comprehend and focus on the context, ensuring that the intended emotional reactions can be effectively elicited.

For the first design goal, we implement a VR video player with a curved display to replicate the natural curvature of human vision. This design enhances immersion by aligning the rendered display with the way participants naturally perceive their surroundings. By simulating this curvature, the viewing experience becomes more engaging and comfortable, reducing visual strain and increasing the sense of presence in immersive environments [59]. This alignment with human visual perception improves spatial awareness and allow a deeper emotional connection with the content, even without the use of panoramic videos. Figure 1 shows a snapshot of a participant's view with non-panoramic video stimulus rendered in the virtual environment.

Second, to improve participants' comprehension of the video content, the display dimensions are adjusted to fit within the mid-peripheral field of vision, which spans approximately 120° horizontally and 60° vertically [58]. This design enables participants to view the entire video frame and easily shift their focus with their eyes, without the need for constant head movement. Our initial testing revealed that exceeding this mid-peripheral field of vision can cause excessive head movement, which significantly distracts the participants from the video and hinders their overall comprehension (and thus, the effectiveness in eliciting the emotion). This issue is largely

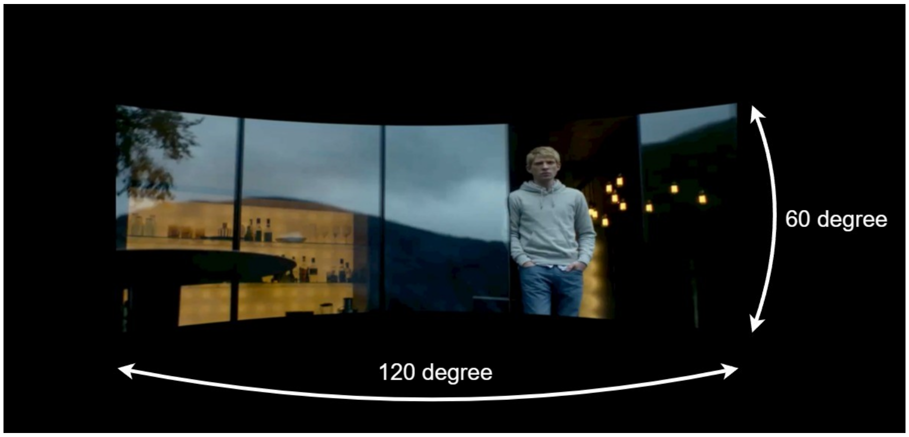


Fig. 1. An illustration of the participant's view when a non-panoramic 2D video is rendered in the virtual environment. The video is displayed on a curved screen to enhance immersion and provide a more comfortable viewing experience. To avoid occupying the entire field of view, the video dimensions are adjusted to fit within the participant's mid-peripheral field of vision (120° horizontally and 60° vertically). This design significantly reduces distractions from unnecessary head movements during data collection, helping participants maintain focus on the video content.

eliminated after adjusting the display size to fit within the mid-peripheral field of vision. We implemented the virtual environment using the Unity engine to ensure compatibility with a wide range of VR hardware.

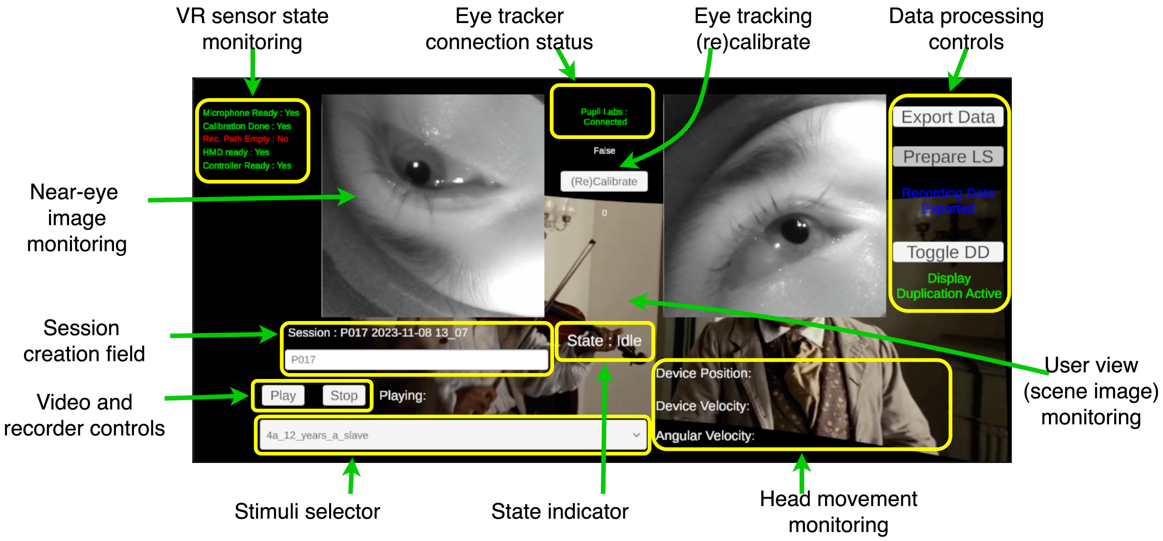
3.2.2 Data Recording Module. The data recording module includes the Recording User Interface and the Data Recording Program.

Recording User Interface (UI). To streamline and ensure the effectiveness of data collection, we design the Recording UI to allow the researcher full control, monitoring, and management of the data collection process. Figure 2 shows the Recording UI in a data collection session. In addition to ensuring reliable data synchronization and recording, a key design priority is that all the collected signals must be continuously monitored and clearly visible to the researcher, enabling quick identification and resolution of any malfunction or issue during the data collection. As shown, the interface integrates multiple functional elements, including control commands (e.g., session creation, eye tracker calibration, data recording, and processing) and monitoring functions (e.g., VR sensors status, eye tracker status, eye images, and head movement tracking).

Data Recording Program. The data recording program is implemented in Unity, with the primary role of handling session control actions initiated by the researcher in the Recording UI and coordinating data flows during recording to ensure synchronized collection of multi-modal signals from various sources. Specifically, it manages signals from the Pupil Labs eye tracker (e.g., near-eye images) and the VR headset (e.g., IMU, scene images, and audio), while coordinating control commands and data flows between the Recording UI, the Virtual Environment, the external eye-tracking tool (Pupil Capture), and the data storage system.

Figure 15 (in Appendix C.2) illustrates how the Data Recording Program manages different data flows from the hardware platforms, i.e., VR headset and the Pupil Labs eye tracker, and the control command from the Recording UI. Four types of signals are recorded and stored in external data storage: audio and IMU data from the VR headset's onboard sensors; near-eye images from the Pupil-Labs eye tracker; and scene images generated in Unity that represent the participant's view in the virtual environment when watching the rendered stimuli.

3.2.3 Data Labeling User Interface. To ensure the label is created as accurate as possible, we perform data annotation right after the participant views each video stimulus. Additionally, instead of applying a single



Microphone Ready : Yes
Calibration Done : Yes
Rec. Path Empty : No
HMD ready : Yes
Controller Ready : Yes

Pupil Labs : Connected

False

0

Recording Data Exported

Display Duplication Active

Session : P017 2023-11-08 13_07

P017

State : Idle

Playing:

4a_12_years_a_slave

Device Position:

Device Velocity:

Angular Velocity:

VR sensor state monitoring
Eye tracker connection status
Eye tracking (re)calibrate
Data processing controls
Near-eye image monitoring
Session creation field
Video and recorder controls
User view (scene image) monitoring
Stimuli selector
State indicator
Head movement monitoring
(Re)Calibrate
Export Data
Prepare LS
Toggle DD
Play
Stop

Fig. 2. An illustration of the Recording User Interface, which offers a range of control and monitoring functions, allowing the researcher to fully operate, monitor, and manage the data collection process. During data collection sessions, the Recording UI is visible only to the researcher and is exclusively used by them.

emotion label to the entire recording, participants are asked to review the video they have just seen and create as many segments as they wish within the recording. The motivation behind is that emotional responses are not uniform throughout the entire duration of the video (in terms of emotional state and intensity). In fact, an effective emotional trigger often requires a buildup of background and context [95, 114]. Depending on how participants interpret the video content, their emotional responses can vary significantly across different phases of the video, with the true emotional peak typically occurring after sufficient context and buildup. Thus, labeling the entire session with a single emotion would overlook these natural fluctuations.

In this work, we design and implement a Data Labeling User Interface based on Label Studio[1] to enable segment-based annotations, allowing us to capture nuanced shifts in emotion type and intensity. Figure 16 (in Appendix C.2) shows a snapshot of the Data Labeling UI. The participant collaborates with the researcher to create and label data segments after viewing each video stimulus. Note that we do not load every data modality into the UI, as they have been synchronized by their timestamps. Instead, only the stimulus video is loaded, with its timestamps serving as a reference. After reviewing and reflecting on the stimuli video, participants can then create data segments and assign labels and intensity rating directly along the video timeline.

4 DATA COLLECTION AND PRELIMINARY VALIDATION

In this section, we begin by introducing the emotion model and video stimuli used in the data collection process, followed by a description of the data collection setup and procedure. We then present the characteristics of the collected dataset and provide a preliminary validation of its effectiveness.

4.1 Emotion Model and Stimuli for Emotion Elicitation

4.1.1 Emotion Model. We consider Ekman's basic emotions [29], i.e., *Happiness, Sadness, Fear, Disgust, Anger, Surprise*, and *Neutral*, in this study. This discrete categorization simplifies the labeling of subjects' emotional

[1] Label Studio is an open-source data labeling platform: https://labelstud.io/

responses. Additionally, to account for the subjectivity and variability in emotions experienced by different participants, we incorporate an emotional intensity rating [108, 138] to capture intensity levels. Each of the discrete emotions is assigned a numerical rating to represent the intensity experienced by participants. This intensity rating allows us to go beyond basic categorization, offering a more nuanced and accurate representation of human emotions in terms of both state and intensity.

4.1.2 *Selection of Emotion Stimuli.*

Emotion stimuli are crucial for creating a high-quality emotion recognition dataset, as they directly affect its utility. In selecting stimuli for data collection, we prioritize those capable of evoking the seven basic emotions, and whose effectiveness has have been validated through large-scale psychological studies.

For emotion elicitation in VR, stimuli are typically categorized as either active or passive [77]. Active stimuli require participants to interact with the virtual environment by engaging with elements or completing tasks [1, 15, 24], whereas passive stimuli involve simply observing content without interaction [63, 70, 98, 111]. Active stimuli include VR games and custom-designed Virtual Reality Environments (VREs) intended to provoke specific emotions. VR games can elicit authentic emotional responses due to their immersive nature [17, 24], but their unpredictability complicates emotional trigger identification and data labeling. In contrast, custom-designed VREs offer a controlled setting where emotional responses can be reliably anticipated at specific moments [1, 15], facilitating structured data collection. However, the development of comprehensive VREs capable of evoking all basic emotions remains in its early stages [68], and no publicly available VRE stimuli meeting our criteria were available at the time of this study. Passive VR stimuli include 360° panoramic videos, standard 2D videos, and images. While 360° videos offer a more immersive experience, most existing datasets focus on valence and arousal [51, 63, 69, 71, 111], and do not support elicitation of all discrete emotions defined in Ekman's model.

To this end, we select non-panoramic 2D videos as stimuli, without compromising on effectiveness. In fact, these videos, often drawn from films, are well-documented in psychological literature [38, 41] and have been proven to effectively elicit targeted basic emotions since as early as 1993 [86]. Specifically, we select movie clips from the database compiled by Zupan et al. [138], which re-validates and expands upon a selection of well-studied film clips [35, 41, 99, 137]. Zupan et al. have applied additional selection criteria to ensure that all clips are of high audio-visual quality, resonate with modern social contexts, cover a wide range of emotions, are in English, are under three minutes to prevent mental fatigue, and are easy to understand without prior knowledge of the storyline or characters.

In our study, data collection was conducted in two stages, each involving a separate set of 14 video clips selected to evoke the seven basic emotions, with two clips per emotion. The detailed information of the selected video clips is provided in Tables 9 and 10 in the Appendix, corresponding to Study One and Study Two, respectively. In both stages, the video clips were chosen from the original database provided by Zupan et al. [138], based on emotional intensity ratings collected from 113 participants. Emotional intensity was rated on a scale from 1 (not at all) to 9 (extremely). For each emotion, we selected the four clips with the highest average emotional intensity. This selection process ensured strong emotional elicitation while maintaining diversity in content and scenarios. Each clip was accompanied by a brief content sentence to help participants contextualize the scene before viewing.

4.2 Data Collection Setup and Procedure

The data collection is conducted in a controlled lab environment, i.e. a quiet, isolated room free from external disturbances and interference, as shown in Figure 3. The room is featured an ergonomic chair and maintains a comfortable temperature ensure participants feel at ease.

The data collection process begins with the participants completing a questionnaire to gather demographic information and look for any conditions that can cause discomfort. Specifically, we exclude participants who

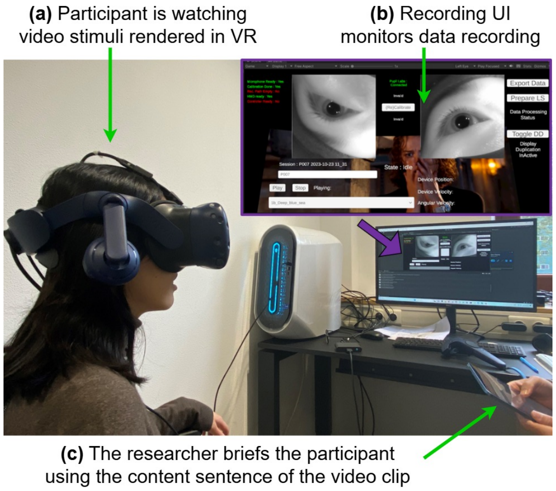

(a) Participant is watching video stimuli rendered in VR
(b) Recording UI monitors data recording
Export Data
Prepare LS
Data Processing Status
Toggle DD
Display Duplication InActive
(c) The researcher briefs the participant using the content sentence of the video clip

Fig. 3. The data collection setup: (a) a participant watching video stimuli displayed in the VR headset, and (b) the researcher briefing the participant using the content sentence of the video clip, and (c) the developed system recording and monitoring the entire data collection process.

(a) Participants demographics

Total number	26
Gender	Female (11), Male (15)
Age	Min: 20, Max: 41, Mean: 26.2 ± 4.2
Ethnicity	Caucasian (11), East Asian (7), Middle Eastern (2), South Asian (4), African (1), Southeast Asian (1)

(b) Participants self-reported states

Stress level	Mean: 3.9, Std: 1.8, CI: (3.2, 4.6) (1: Very Low, 10: Very High)
Fatigue level	Mean: 4.6, Std: 1.8, CI: (3.9, 5.3) (1: Not Fatigued at all, 10: Extremely Fatigued)
Comfort level	Mean: 7.7, Std: 1.6, CI: (7.1, 8.4) (1: Very Uncomfortable, 10: Very Comfortable)

Table 1. (a) Demographic information of the 26 participants shows a diverse representation in gender and ethnicity. (b) The statistics of self-reported states, including mean, standard deviation, and 95% confidence interval, indicates low stress and fatigue levels among participants prior to data collection, with participants expressing comfort with the experiment setup.

experience motion sickness in VR, as well as those with epilepsy, anxiety, or claustrophobia, due to potential increased symptoms or discomfort in VR environments. We measure each participant's inter-pupillary distance to adjust the VR headset for a comfortable and accurate fit. Once the VR headset is in place, the researcher checks the data streams, activates SteamVR in Night Mode to minimize visual distractions, and calibrates the participant's gaze in Unity. The VR headset audio is also set to a comfortable level. Participants report their levels of stress, fatigue, and comfort before the data collection process. For each session, we follow the screen marker calibration choreography provided by Pupil Labs. Specifically, we developed a Unity application that displays the screen marker within the virtual environment and sends the calibration data to Pupil Capture upon completion. As shown previously in Figure 2, our data recording program monitors whether calibration has been completed before each data collection session, and the user interface allows researchers to easily re-calibrate when needed.

A total of 26 participants contributed to our dataset through two stages of data collection. In Study A, we recruited 20 participants who completed the first round of data collection. For Study B, an additional six participants were invited to take part in a second round of data collection using a new set of video stimuli. The demographic information and self-reported emotional state statistics are presented in Tables 1 (a) and (b), respectively. Our dataset includes a diverse sample in terms of gender and ethnicity. Self-reported data from the participants indicate generally low levels of stress and fatigue, along with high levels of comfort throughout the study. Participants for the second stage were newly recruited to avoid bias from prior exposure to similar stimuli in the first stage.

Stimuli Presentation and Playback. During data collection, the 14 video stimuli are played in a predefined sequence. Playing a high-arousal clip (e.g., video evokes fear) immediately before or after a low-arousal clip (e.g., video evokes happiness) could cause unintended emotional carryover effects [136]. According to the excitation-transfer theory, residual physiological arousal from one event can intensify emotional reactions to a following event, even if they're unrelated. For instance, a recent study has found that participants who watched high-arousal negative video clips experienced retrograde memory impairments for neutral information presented afterward [105]. To address this, we establish a specific order: starting with Neutral (a), followed by Surprise (a), then Neutral (b), Surprise (b), and continuing through Happiness (a,b), Sadness (a,b), Anger (a,b), Disgust (a,b), and Fear (a,b). Here, 'a' and 'b' denote the two clips intended to evoke the same emotional response, as outlined in Table 9. Moreover, participants are given sufficient break time between each video and engaged in labeling the most recently viewed video to help minimize emotional carryover effects.

To introduce an element of unpredictability, the initial Surprise (a) clip is presented to the participants as Neutral, while the Surprise (b) clip is labeled as Happiness. Additionally, since these video clips are all extracted from movies and lack sufficient standalone context, before each video stimulus is played, we provide participants with a brief overview by reading a context sentence provided by the original database to prepare them emotionally without diluting the intended effect. The corresponding context sentences for each video clip are listed in Table 9. This briefing has proven effective in enhancing participants' understanding of the video content, and thereby improving the effectiveness of the stimuli in eliciting the intended emotional responses [138]. For clips designed to evoke surprise, context sentences were slightly modified to maintain the element of surprise.

Data Labeling. After each video, the participant and researcher use the Data Labeling UI together (shown in Figure 16) to identify and annotate emotional segments in the recordings. In addition to assigning emotion labels to each segment, participants are asked to rate the intensity of the labeled emotion on a scale from 1 to 10, reflecting their emotional response intensity [63]. Specifically, participants answer the question: "How [happy, sad, angry, etc.] would you say you felt during this segment of the video on a scale from 1 to 10?" A rating of 1 indicates no experience of the labeled emotion, while a rating of 10 reflects a very intense experience. A rating of 6 represents a noticeable experience of the emotion with minimal intensity.

4.3 Dataset Characteristics

An overview of our dataset is provided in Table 2. The dataset includes a diverse set of data types, i.e., eye movement [111], head movement [97, 121], and periocular images [91], each known to contribute to emotion recognition. As discussed in Section 3.2.3, we consider data segments created and labeled by the participants as effective emotional periods, rather than using the entire recording. Thus, the final length of these segmented data is shorter than the total duration of the emotional stimuli presented in Tables 9 and 10. In the end, as shown in Table 2 the segmented data length for each of the seven emotion types ranges from 279 seconds (Surprised) to 4,067 seconds (Happiness) for Study A, and from 121 seconds (Surprised) to 900 seconds (Anger) for Study B. We also release the full data recordings with our dataset for future comparison studies and research purposes.

Multi-faceted, High-resolution Eye-tracking data. Our dataset includes high-resolution eye-tracking data captured at a high sampling rate: two-dimensional gaze direction is recorded at 240Hz, while pupil diameter and pupil position for both eyes are tracked at 120Hz. This sampling rate, and thus the data resolution, is four times higher than that of the current state-of-the-art VREED dataset [111] (240Hz vs. 60Hz). This enhanced eye-tracking resolution enables precise analysis of subtle gaze shifts and pupil dynamics during emotional fluctuations, offering valuable insights into moment-to-moment emotional and cognitive processes.

Additionally, our dataset includes near-eye grayscale images recorded at 120fps with a 400×400 resolution, capturing fine details in the periocular regions during various emotional states. Periocular regions, encompassing the eyelids, eyebrows, and surrounding skin, are crucial for emotion recognition [91], as they reveal rich visual cues related to emotional responses. With such a high frame rate, our data captures subtle movements in this area, such as micro-expressions, eyebrow shifts, and muscle tension, to effectively reflect underlying emotions. This newly introduced periocular video data provides a valuable resource for emotion analysis in VR and head-mounted display contexts. The VR scene recordings, representing the participant's view within the virtual environment, are

Table 2. Summary of the collected dataset. Our dataset offers high-resolution, multi-faceted eye-tracking data, capturing 2D gaze direction, pupil diameter, and pupil position at four times the resolution of existing state-of-the-art datasets. Additionally, it includes periocular video, which records fine-grained motions such as micro-expressions and eyebrow movements—features highly relevant for emotion recognition.

Recording length (Study A)	
Total length (in second) of segmented data for each emotion	Neutral (1,477 s), Surprised (279 s), Happiness (4,067 s), Sadness (1,885 s), Anger (2,694 s), Disgust (2,443 s), and Fear (3,142 s)
Total length (in second) of full recordings for each emotion	Neutral (1,477 s), Surprised (2,640 s), Happiness (5,240 s), Sadness (4,960 s), Anger (4,680 s), Disgust (3,900 s), and Fear (4,040 s)
Recording length (Study B)	
Total length (in second) of segmented data for each emotion	Neutral (807 s), Surprised (121 s), Happiness (422 s), Sadness (877 s), Anger (900 s), Disgust (403 s), and Fear (879 s)
Total length (in second) of full recordings for each emotion	Neutral (807 s), Surprised (348 s), Happiness (960 s), Sadness (1,248 s), Anger (1,284 s), Disgust (810 s), and Fear (1,626 s)
Video stimuli	
Emotion intensity rating	Participant self-reported emotional response intensity ranges.
Stimuli	Each emotion is elicited by a pair of stimuli ('a' and 'b') as listed in Tables 9 and 10 shown in Appendix. For both studies, two data sessions, each with seven video stimuli, are presented to each participant.
Eye tracking data	
Gaze coordinates	2-dimensional gaze direction, with 240Hz sampling rate.
Pupil diameter	1-dimensional data, for both left and right eyes, with 120Hz sampling rate.
Pupil position	2-dimensional data indicates pupil position in the near-eye images, for both left and right eyes, with 120Hz sampling rate.
Periocular video recording	Grayscale video recorded by two near-eye cameras for both left and right eyes, with 120fps frame rate and 400 × 400 resolution.
VR scene recording	
Scene recording	Participant's view in the virtual environment (as shown in Figure 1) captured at 120fps frame rate with 648 × 480 resolution.
Head movement	
9-DoF IMU measurements	9-dimensional time-series data capturing angular velocity, linear acceleration, and orientation at 30Hz sampling rate.

collected at 120fps with a resolution of 648×480. Additionally, head movement data are recorded through 9-DoF IMU sensors, providing a 9-dimensional dataset (including angular velocity, linear acceleration, and orientation) sampled at 30Hz. We believe this new multi-modal dataset (eye tracking, scene view, and head movement), along with multi-faceted eye-tracking signals (gaze direction, pupil dynamics, and periocular imagery) offers a rich resource for advancing future research in emotion recognition within VR environments. Applications and use cases of our dataset are further discussed in Section 6.

Ethics. The research received approval from the Research Ethics Review Group at the host institution, adhering strictly to all institutional guidelines. All participants sign a consent form before participating and are informed of their right to withdraw from the study at any time. To address privacy and security concerns associated with physiological data, all identifying information are removed from the dataset to ensure participant anonymity. Additionally, all data will be securely stored in a local data repository. Any future requests for dataset access will be thoroughly reviewed, and all data downloads will be logged to maintain an access record.

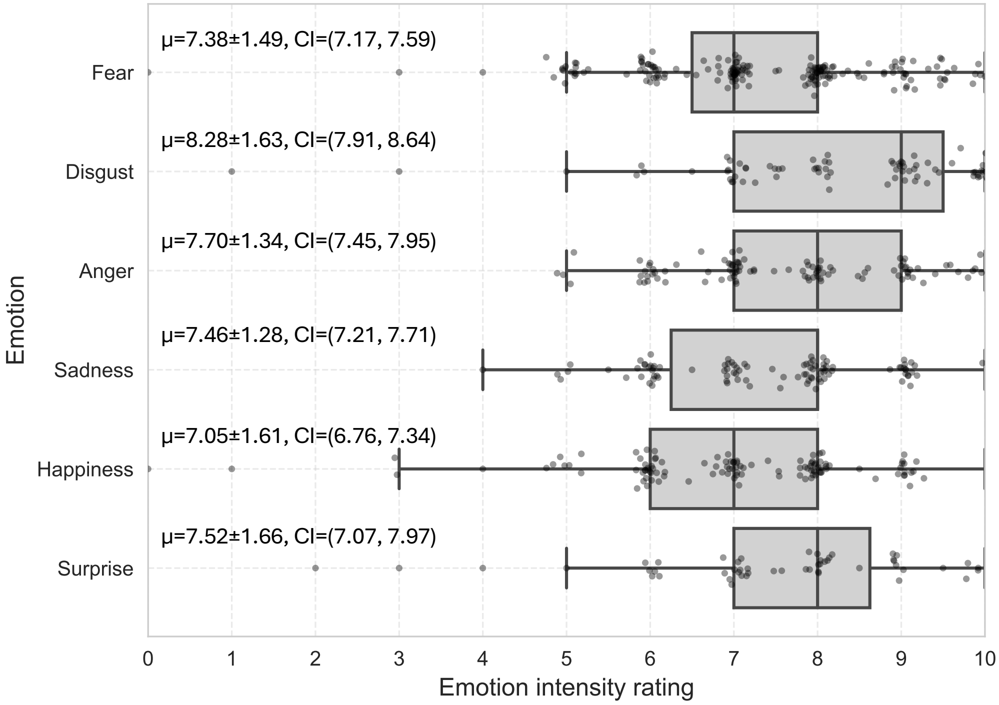


Fig. 4. Participants reported emotional intensity ratings for six emotions, displayed in a boxplot with scatter points. Each dot represents a participant's rating for an emotion segment they created and labeled. There are 189, 77, 107, 99, 117, and 52 segments for Fear, Disgust, Anger, Sadness, Happiness, and Surprise, respectively. For each emotion, we also display the mean (μ) and the 95% confidence interval of the intensity ratings. Overall, our experiment effectively evoked the intended emotions in participants, achieving high intensity levels across emotions.

4.4 Emotion Elicitation Measurement

Below, we analyze participants' self-reported emotion intensity ratings to assess how effectively our experiment (i.e., the combination of selected stimuli and VR rendering) elicited the intended emotions. We use an emotional response rating scale similar to that of Zupan et al. [137] to gauge emotion elicitation. Participants rated the intensity of each labeled emotion on a scale from 1 to 10, where a rating of 1 indicates no experience of the emotion, 10 reflects a very intense experience, and 6 suggests a noticeable but moderate intensity.

In total, we received 641 emotion intensity ratings across 189, 77, 107, 99, 117, and 52 participant-created segments for the emotions Fear, Disgust, Anger, Sadness, Happiness, and Surprise, respectively. Figure 4 presents a boxplot with scatter points, where each dot represents a participant's rating for a segment they created and annotated. For each emotion, we also provide the mean (μ) and 95% confidence interval (CI) for the intensity ratings, indicating a notably high success rate in emotion elicitation.

5 VALIDATION

In this section, we present preliminary benchmarks to showcase the potential of our dataset in emotion recognition. We emphasize the use of the new periocular video data collected in our dataset, and combined it with the widely used eye movement time-series signals [43, 111, 122], to enhance recognition accuracy. We compare the effectiveness of using periocular data alone, eye movement data alone, and their combination to evaluate each method's contribution to overall emotion recognition performance. In addition to the classification results, the Appendix presents a case study that qualitatively and quantitatively analyzes pupil diameter variations in response to emotional changes. This highlights the potential of the proposed dataset for future research on emotion-induced physiological responses and eye movement-based biometric feature analysis.

5.1 Multi-faceted Eye Tracking Fusion

Overall design. To effectively leverage both periocular eye videos and eye movement time-series signals (gaze direction and pupil diameter) for emotion recognition, we propose the multi-faceted eye-tracking fusion method. A naive approach for data fusion would be to feed all available signals into a complex neural network for feature extraction and classification. However, for VR-based emotion recognition, we must consider the model's computational complexity. Given the resource constraints of VR platforms, feeding high-dimensional, large-scale data into a deep learning model can lead to significant computational latency. In particular, the original periocular eye videos in our dataset are recorded at a high frame rate of 120fps. Directly using these high-frame-rate videos as neural network inputs can significantly increase computational costs due to the large data volume. By contrast, the eye movement time-series signals are more computationally efficient because they are lower-dimensional. However, this efficiency comes at the cost of information richness. Eye movement time-series signals capture quantitative changes in the time domain, e.g., changes in gaze direction, fixation, or pupil dynamics, but lacks the spatial context and fine-grained detail that the periocular eye video frames can provide, e.g., changes in eye shape, wrinkles, and subtle movements around the eyes that can effectively signal emotional states.

To balance the trade-off between model performance in emotion recognition and computational efficiency, we utilize eye movement time-series signals at their maximum available sampling rates, i.e., 240Hz for gaze direction and 120Hz for pupil diameter, while down-sampling the periocular eye videos to a lower frame rate (e.g., 10fps). This multi-faceted fusion approach retains rich information while minimizing computational demands, creating a practical and efficient solution for emotion recognition in VR contexts. Below, We refer to this method as the **multi-faceted** eye tracking fusion. As shown in Figure 5, the proposed model comprises four main components: a video feature extractor, an eye movement feature extractor, a feature fusion module, and an emotion classifier. The video feature extractor learns features from the periocular videos for both the left and right eyes, while the eye movement feature extractor captures features from eye movement data. The feature fusion module then combines the extracted features from these different input sources. Finally, the emotion classifier uses the fused features to predict the user's emotional state. We present the details for each component below.

Periocular Feature Extractor. We adopt the Video Vision Transformer (ViViT) [4] as the periocular feature extractor, denoted as $\mathbf{F}_V$, to capture spatio-temporal features from the recorded eye region frames. Given an input periocular video clip $V \in \mathbb{R}^{T\times W\times H}$ consisting of T gray-scale frames with width W and height H, the extractor $\mathbf{F}_V$ maps the clip to a video feature vector $f_v \in \mathbb{R}^d$, expressed as $f_v = \mathbf{F}_V(V)$. We can set the T to be a low number to reduce the computational cost. Note that, We process the periocular videos from the left and right eyes using two periocular feature extractors with shared model weights. Moreover, we retain information from both eyes, rather than selecting only one of them, to account for the potential feature differences introduced by the dominant eye phenomenon [115]. In terms of model design, the periocular feature extractor is designed with a hidden dimension of 256 and includes three spatial transformer layers followed by a single temporal layer, each with eight attention heads. To improve generalization, we apply a dropout rate of 0.1 in both the encoder and embedding layers.

Eye movement Feature Extractor. We design the eye movement feature extractor $\mathbf{F}_E$ based on the multivariate time-series transformer framework [127]. We represent a sequence of eye movement time-series signals as $X \in \mathbb{R}^{N\times M}$, where N is the sequence length and M is the dimensionality. In our model, we concatenate the 1D pupil diameter for both left and right eyes and the 2D gaze directions as the inputs, resulting a four-dimensional

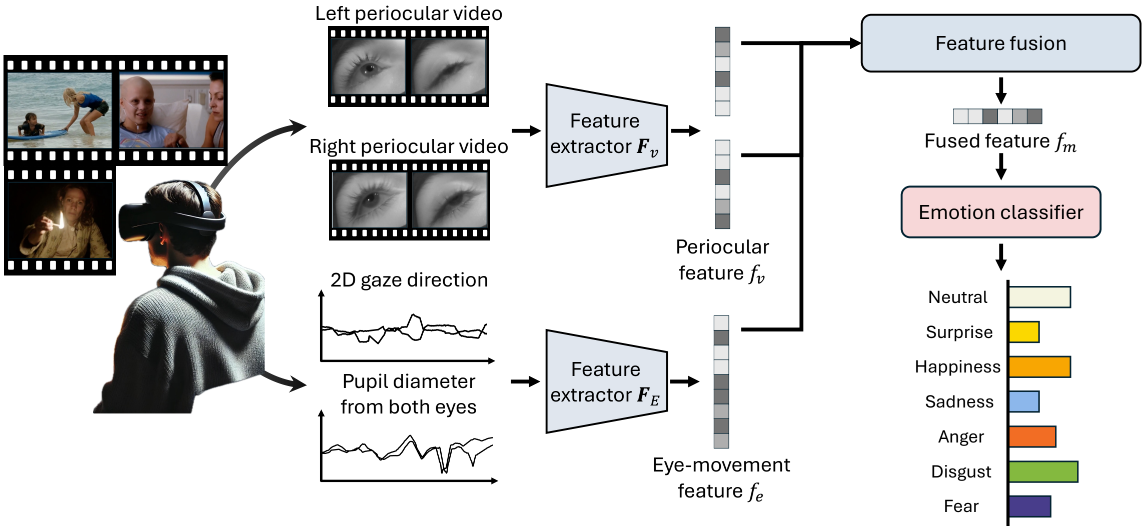


Fig. 5. Overview of the proposed multi-faceted emotion recognition model, which extracts features from both periocular eye videos and eye movement data, and then fused them for emotion recognition.

eye movement time-series signal X. The feature extractor $\mathbf{F}_E$ then extracts a k-dimensional feature vector $f_e \in \mathbb{R}^k$ from X, i.e., $f_e = \mathbf{F}_E(X)$. It employs a model dimension of 64 and uses eight attention heads within a single transformer layer. Like the ViViT module, it includes a dropout rate of 0.1 for regularization.

Multi-modal Feature Fusion. With the extracted features from periocular videos and eye movement time-series signals, we design a feature-fusion module to combine them for the emotion recognition task. First, we linearly project the features extracted from the left periocular video, right periocular video, and eye movement time-series signals into three separate h-dimensional vectors using distinct fully connected layers. These three projected h-dimensional vectors are then stacked to form a feature matrix of dimension $h \times 3$. This matrix is fed into a transformer encoder to apply self-attention for effective feature fusion [21, 88]. We denote the resulting fused multi-modal feature as f_m.

Emotion Classifier. Finally, we apply a fully connected layer with a softmax function to classify the emotion based on the fused multi-modal feature f_m. The emotion recognition model is trained by minimizing the cross-entropy loss between the predicted emotions and the ground-truth emotion labels. The classification includes Ekman's six basic emotions, along with a neutral state.

5.2 Data Pre-processing

To minimize the noise inherent in eye movement data, both gaze directions and pupil diameters, we pre-process the raw signals with several filters and operations. First, a confidence filter removes corrupted eye movement samples by excluding those with a confidence score below a threshold of 0.6. Next, we calculate the median absolute deviation (MAD) using a one-second window and discard samples with values exceeding three times the MAD. For pupil diameter samples, we apply a Z-score filter [72] with a threshold of 0.1, followed by a low-pass filter (cutoff = 5Hz). These filters smooth out rapid fluctuations in pupil diameter [56, 102], given that pupil diameter changes respond more slowly than gaze movements [109]. Since each filtering step removes some eye movement samples, we resample the data back to its original length using linear interpolation after each filter. Finally, we standardize and normalize the filtered eye movement samples within each session.

For the periocular video data, we first apply horizontal and vertical flips to the right-eye video to align its spatial orientation with the left eye. We then resize each video frame from 400 × 400 to 224 × 224 pixels, balancing computational efficiency with sufficient detail to retain informative features. To standardize pixel values across both videos, we normalize by dividing each pixel by 255, following standard preprocessing practices for deep learning models [39].

For both eye movement time-series and periocular video data, we extract the emotion-elicited segments based on the subject's self-reported emotion intensity rating. Each segment begins when the subject reports an intensity level above six and continues until the end of the video. The periocular recordings are down-sample from 120fps to 10fps. The eye movement time-series is sampled at 120Hz. We create data instances (for modeling training and testing) from these segments using non-overlapping sliding windows with a maximum length of two seconds.

5.3 Experimental Design

5.3.1 Baselines. We compare our proposed **multi-faceted approach** with two baseline methods: deep learning model that uses only periocular video and a model that leverages only the eye movement time-series data.

Specifically, the (1) **periocular method** uses only periocular video recordings as input, employing the upper portion of the architecture shown in Figure 5. This includes feature fusion to combine features extracted from the periocular videos of both eyes. The (2) **eye movement method**, on the other hand, uses only eye movement data, i.e., gaze directions and pupil diameters of both eyes, as input and utilizes the lower portion of the architecture in Figure 5, excluding the feature fusion module.

It is important to note that, because the use of a more advanced deep learning architecture, i.e., the transformer [4, 127], these two methods serve as stronger baselines compared to existing works [42, 43, 111, 122] that rely on conventional machine learning algorithms, e.g., random forest and support vector machine, with hand-crafted features, as deep learning-based methods have been shown to achieve superior classification performance on gaze and eye movement-based inputs [60, 61].

The goal of this comparison is to highlight the effectiveness of combining periocular video and eye movement time-series data for emotion recognition, showcasing the advantages of our fusion strategy. Additionally, it underscores the value of the newly incorporated periocular videos in our dataset, demonstrating how these data enhance emotion recognition research by providing richer, complementary visual cues that improve model performance when fused with traditional eye movement signals.

5.3.2 Experiment Setups. We perform five-fold cross-validation on the 20 subjects, with each fold containing four subjects. In our dataset, as each of the seven emotions contains two data recordings, we treat them as two separate data sessions. We consider the following settings in the evaluation.

First, in the (1) **Pre-train only** setting, we train a randomly initialized model using data from four folds (comprising 16 subjects, each with two sessions of data) and test it on the remaining fold (including 4 subjects, each with two sessions of data). We report the average performance across all five folds to assess the model's ability to generalize to unseen subjects.

Second, in the (2) **fine-tune** setting, we begin with the model initially trained with the pre-train setting and further fine-tune it using 10% of instances sampled from the test set. Here, the test set consists of four unseen subjects, each with two data sessions. Without shuffling the test set, we select instances for fine-tuning from the beginning of each data session for the four subjects, while reserving the remaining instances within the same session for testing. This setting reflects a real-world scenario where only a limited amount of data from the start of a recording, such as the initial interactions with a new user, is available for model adaptation. By fine-tuning on these data instances obtained at the beginning of the VR interaction, we can then assess how well the model generalizes to subsequent interactions within the same session. More specifically, we employ two different strategies to evaluate model performance within the fine-tune setup:

- **Same-session:** We use instances from both sessions for fine-tuning and testing, meaning that data instances for fine-tuning and model testing are coming from the same data sessions, though sequentially without random shuffle. This allows us to evaluate how well the model adapts when it has a small amount of fine-tuning data from all the sessions (subjects and video stimuli) it will encounter in testing.
- **Cross-session:** We fine-tune the model on instances from one session and test it on the other session. Since each subject has two sessions per emotion, this setup is repeated twice for each subject, and we report the averaged performance. The cross-session approach examines the model's ability to generalize across sessions, reflecting cases where fine-tuning data is limited to one session, yet testing occurs on a different set of video stimuli.

Performance metric. For all experiments, we report both accuracy and weighted F1-score as performance metrics. While the accuracy provides a high-level view of the model's overall correctness, the weighted F1-score accounts for class imbalance among the emotion categories.

5.4 Implementation Details

The training process consisted of 25 epochs, using the Adam optimizer with a learning rate scheduler. The scheduler included a five-epoch warm-up phase, gradually increasing to a peak learning rate of $8e-5$. This was followed by 20 epochs of cosine annealing, where the learning rate decreased to $8e-6$. In the final five epochs, the lower learning rate was maintained to ensure stable convergence. We applied a label smoothing factor of 0.4 and consistently allocated 30% of the training data for validation. For the fine-tuning setting, we use the same optimizer and learning rate scheduler as in the initial training but extend each phase duration by five times to support gradual adaptation. The peak learning rate is reduced to $1e-5$, with a final learning rate of $1e-6$ to minimize overfitting. Label smoothing is omitted in this phase.

The training process outlined above applies to both the multi-faceted and periocular methods. For the eye movement method, however, we increase the number of training epochs and use a higher learning rate to ensure convergence. For model configuration, we set the periocular feature extraction window T to five, with video frame width W and height H at 16. The dimensionality of the periocular feature d was set to 1024, while the eye movement feature dimension k was set to 256. We configured the feature fusion module dimension h to 256.

5.5 Evaluation Results

5.5.1 Performance Analysis in Pre-train Settings. We show the performance of different methods in Table 3 with different time window sizes. The periocular only method achieves better performance than the eye movement only method across all window sizes with big margins. The multi-faceted method further outperforms the periocular only method in most settings.

Taking the 2.0-second window size as an example, the multi-faceted approach shows a 15.6% improvement over the periocular only method (0.52 vs 0.45) and a 73.3% improvement over the eye movements-only method (0.52 vs 0.30). However, an exception occurs at the 0.5-second window size, where the multi-faceted approach slightly underperforms (F1-score of 0.43) compared to the periocular-only method (F1-score of 0.45). It could be attributed to the limited information provided by eye movements in such a short time frame, potentially introducing noise rather than valuable features to the multi-faceted approach. This suggests that there exists a minimum input window size to effectively leverage eye movement data for emotion recognition.

Overall, the results underscore the wealth of information contained within the periocular images for emotion recognition. The multi-faceted method, for instance, combines the strengths of both the high-frequency time-domain features from eye movement signal and the fine-grained spatial details captured in the periocular video.

Table 3. Recognition performance of different methods in the pre-train only setting with different window sizes. Both periocular only and multi-faceted methods achieve better performance than the eye movement only approach, underscoring the wealth of information contained within periocular image frames for emotion recognition.

Method	**Metric**	**Window size (s)**			
		0.5	**1.0**	**1.5**	**2.0**
Eye movement	F1-score	0.26	0.27	0.29	0.30
	Accuracy	0.28	0.29	0.31	0.31
Periocular	F1-score	0.45	0.44	0.46	0.45
	Accuracy	0.45	0.44	0.47	0.45
Multi-faceted	F1-score	0.43	0.44	0.46	**0.52**
	Accuracy	0.43	0.44	0.47	**0.52**

Table 4. Recognition performance of different methods in the fine-tuning setting for both same-session and cross-session configurations. The results indicate that the multi-faceted method demonstrates superior generalizability, especially in cross-session configurations where testing is conducted on a different set of unseen video stimuli. This highlights the robustness and capability of the multi-faceted approach in adapting to new and unseen scenarios.

Method	**10% Same-session**		**10% Cross-session**	
	F1-score	**Accuracy**	**F1-score**	**Accuracy**
Eye movement	0.46	0.49	0.23	0.25
Periocular	0.82	0.82	0.54	0.55
Multi-faceted	**0.84**	0.85	**0.70**	0.71

5.5.2 Performance Analysis in Fine-tune Settings. Below, we fix the window size to one second to balance performance and efficiency. Table 4 shows the results of 10% proportional fine-tuning for both same-session and cross-session configurations. The multi-faceted approach achieves the highest F1-score in all examined cases, reaching 0.84 in the same-session configuration and 0.70 in the cross-session configuration. By contrast, the eye movement-only method shows poor cross-session accuracy, with an F1-score of just 0.23, indicating that eye movement data alone is insufficient to capture generalized features that are robust for emotion recognition in unseen video stimuli.

The periocular-only method performs comparably to the multi-faceted method in the same-session setting; however, in the more challenging cross-session setting, where testing is conducted on a different set of unseen video stimuli, the multi-faceted approach shows a clear advantage. This demonstrates the benefit of integrating both eye movement and periocular information for emotion recognition, underscoring the multi-faceted approach's adaptability and robustness in handling new and unseen scenarios.

5.6 Few-shot Performance in Cross-session Configuration

In the previous fine-tuning experiments, we used 10% of the data to fine-tune the pre-trained model, which corresponds to approximately five training instances per class on average. Given that the number of training instances available for fine-tuning is a crucial factor in real-world applications scenarios, we further explore its impact by conducting a few-shot study in the challenging cross-session configuration. Specifically, the pre-trained model is fine-tuned using a few-shot instances from one of the data sessions, and test it on data from the other session (a different set of video stimuli).

Table 5. Performance when a few-shot of instances are used for the fine-tuning in the cross-session configuration. The multi-faceted approach consistently and significantly outperforms the periocular-only baseline across all few-shot settings, highlighting its strong generalization capability in real-world scenarios where labeled training samples from unseen scenarios (new video stimuli) are difficult to obtain.

Method	**Metric**	**1-shot**	**2-shot**	**3-shot***	**4-shot***	5-shot*
Periocular	F1-score	0.53	0.56	0.57	0.57	0.57
	Accuracy	0.53	0.56	0.57	0.57	0.57
Multi-faceted	F1-score	0.67	0.68	0.69	0.70	0.70
	Accuracy	0.66	0.67	0.69	0.69	0.70

*: For emotion class *Surprise*, at most two shots (i.e., two training instances) are taken.

Table 5 compares the recognition performance of the periocular-only and multi-faceted methods across different few-shot settings, where one to five instances per class are used for fine-tuning. For the emotion class *Surprise*, we limit the number of instances to a maximum of two due to the scarcity of data for this class. We also exclude the results for the eye movement-only method, as its performance remains poor given the limited information that eye movement data alone can provide (as demonstrated in Table 4).

The results indicate that the multi-faceted approach consistently and significantly outperforms the periocular-only baseline across all few-shot settings. With just one shot, the multi-faceted method achieves an F1-score of 0.67, compared to 0.53 for the periocular baseline. This performance gap persists as the number of shots increases; at the 4-shot setting, multi-faceted method reaches an F1-score of 0.70, while the periocular-only method peaks at 0.57. These findings highlight the multi-faceted method's strong generalization capability with limited training data and demonstrate its robustness in real-world scenarios where labeled training samples from unseen scenarios (new video stimuli) are difficult to obtain.

5.7 Impact of Eye Movement Signal Window Size

In the previous experiments, we use short window sizes, i.e., up to two seconds, for the eye movement-only method. This choice was driven by the high computational demands of the video feature extractor when processing longer periocular videos. However, prior emotion recognition studies using eye movement time-series data have explored considerably longer window sizes, such as 10 seconds [53] and even 180 seconds [111]. To ensure a fair comparison and gain a deeper understanding of how window size impacts the performance of the eye movement baseline, we evaluate its performance across a range of time window sizes up to 15 seconds. Note that extending the time window beyond 15 seconds was not feasible in this study, as the natural occurrence of the seven emotions we investigate does not typically persist for longer durations.

As shown in Table 6, there is a consistent trend of improvement as the time window size increases. In the pre-train setting, the F1-score rises from 0.26 at 0.5 seconds to 0.39 at 15 seconds. A similar pattern is observed for fine-tuning: for the 10% same-session scenario, the F1-score improves from 0.46 with a 1-second window to 0.54 with a 15-second window, a 17.4% increase. This suggests that a larger window size continue to capture additional informative features beneficial for emotion recognition.

However, it is important to note that even with a 15-second window, performance (0.39 in the pre-train setting, 0.54 in the fine-tune setting) remains lower than that of the multi-faceted method, which achieves 0.44 in the pre-train setting and 0.84 in the fine-tune setting using shorter, 1-second windows (as reported in Tables 3 and 4). These results indicate that while longer window sizes provide additional information for emotion recognition when using the eye movement baseline, the performance ceiling of this approach is limited. The multi-faceted approach

Table 6. Impact of window size on the eye movement-only method. We examine the effect of various window sizes on performance for both pre-train only and fine-tune settings.

Method	Setting	Window size (s)						
		0.5	1	1.5	2	5	10	15
Eye movement	F1-score (Pre-train)	0.26	0.27	0.29	0.29	0.34	0.37	0.39
	Accuracy (Pre-train)	0.28	0.29	0.31	0.31	0.34	0.38	0.40
Eye movement*	F1-score (Fine-tune)	0.46	0.46	0.44	0.43	0.50	0.52	0.54
	Accuracy (Fine-tune)	0.49	0.49	0.48	0.46	0.51	0.53	0.56

*: 10% same-session fine-tune.

demonstrates superior performance, highlighting the effectiveness of fusion high-frequency eye movement data with periocular video for emotion recognition.

5.8 Impact of Periocular Video Frame Rate

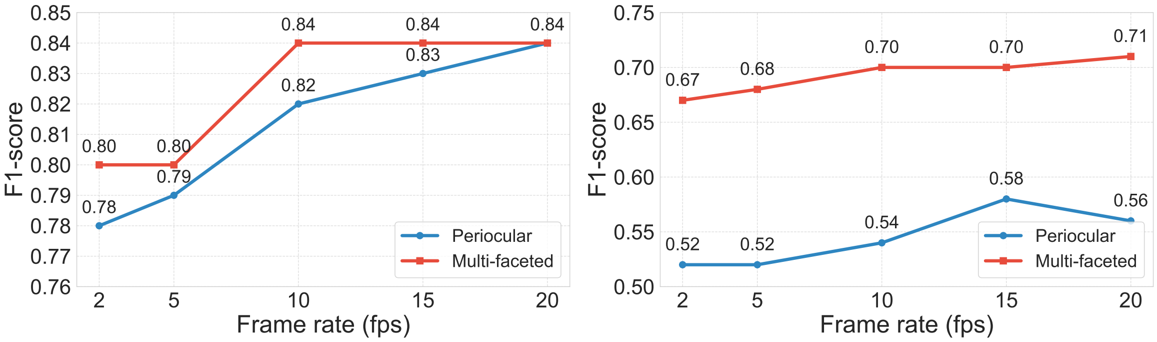
Although our dataset contains periocular video at up to 120fps, it is not resource-efficient to feed all video data into the neural network due to computational constraints on consumer devices. Therefore, in the previous experiments, we down-sample the periocular recordings to 10fps while maintaining high-frequency eye movement time-series data at 120Hz to compensate for the reduced temporal resolution of periocular videos. However, higher frame rate periocular video should theoretically help in improving performance. To test this hypothesis, we evaluate the impact of varying the frame rate of periocular videos while keeping eye movement time-series data at 120Hz. We set the window size to one second.

Figure 6 shows results for fine-tune setting across different frame rates. These results show that a higher frame rate generally leads to a better performance for both the periocular-only baseline and the multi-faceted method. Specifically, in the same-session scenario, the performance difference between the periocular-only and multi-faceted methods is minimal. However, in the more practical and challenging cross-session scenario, the multi-faceted method shows substantial improvements over the periocular-only method across all examined frame rates. For example, at 20fps, the multi-faceted method achieves an F1-score of 0.71, representing a 26.8% improvement over the periocular baseline's 0.56. We also notice that the performance of the multi-faceted method become stable when the frame rate increases to 15fps. We believe this is due to the current lightweight design of the periocular feature extractor: with more frames are added, the periocular video content reaches a point of saturation. Consequently, the feature extractor's ability to capture meaningful features does not improve significantly with the increased input load, as it becomes less efficient in discerning new information from the additional data.

Overall, these results validate the approach of combining low frame rate periocular video with high-frequency eye movement time-series data. The consistency of the multi-faceted method outperforming across all frame rates demonstrates its robustness and effectiveness in emotion recognition tasks.

5.9 Fine-tuning with Empirically Estimated Labels

In previous experiments, we use the self-reported annotations from subjects to crop and label the raw data recordings. Specifically, participants rate the intensity of their emotions for each segment they created. In evaluation, we consider an emotion has been effectively elicited if a segment received an intensity rating of six or higher. We then mark the first segment with an intensity rating of six as the starting point of the emotion and crop the raw data recording from this segment onward. However, in real-world scenarios, user-defined segments



(a) Performance comparison in same-session setting.

(b) Performance comparison in cross-session setting.

Fig. 6. Recognition performance of the multi-faceted method across different frame rates for periocular video, evaluated in the fine-tuning setting for both same-session and cross-session scenarios. The y-axis shows the F1-score for emotion recognition (higher is better), while the x-axis represents various frame rates, ranging from two to 20fps.

Table 7. Recognition performance with statistically derived labels. We evaluate the periocular and multi-faceted methods in the fine-tune setting using data labeled through statistically derived estimates.

Method	**10% Same-session**		**10% Cross-session**	
	F1-score	**Accuracy**	**F1-score**	**Accuracy**
Periocular	0.83	0.83	0.54	0.56
Multi-faceted	0.85	0.85	0.71	0.72

with emotion intensity rating are unavailable. Even during a calibration process, when an emotion-aware system might feed video stimuli designed to elicit specific emotions to the user for data collection, it would be impractical to also ask users to manually mark and annotate their emotional segments as was done during our data collection.

To simulate a more practical scenario, we use leave-one-subject-out cross-validation. We assume that detailed labels and annotations are available from 19 subjects, and assume the video stimuli can reliably elicit certain emotions. Using data from these 19 subjects, we then empirically estimate the point in the stimuli where the emotion of the unseen subject is likely to be triggered. Specifically, we set the estimated emotion onset to the lower limit of the 95% confidence interval derived from the starting points of the 19 subjects. This method enables us to approximate emotion onset times without relying on individual self-reports from the new subject.

Table 7 presents the recognition performance of the periocular only and multi-faceted methods when using the estimated labels for both training and testing. Both models achieve reasonable performance in same-session and cross-session settings. Moreover, as these estimated labels are statistically derived from the self-reported data of the participants, they demonstrate the reliability of the labels in our dataset: the statistically obtained annotations can still be used to train the machine learning model and achieve good classification performance on new users even when personalized data annotation is unavailable.

Figures 7 and 8 (in Appendix B) give details of the confusion matrices of the recognition results for Subjects 3 and 7 as examples. Note that the cross-session results in Table 7 represent the average performance across the two cross-session A and B. In the *same-session setting*, both the multi-faceted and periocular-based methods achieve high accuracy for emotions such as happiness, sadness, and anger. However, both approaches face challenges in

Table 8. Performance when a few-shot of instances are used for the fine-tuning in the cross-study configuration. The multi-faceted approach consistently and significantly outperforms the other baselines across all few-shot settings, highlighting its strong generalization capability in real-world scenarios where labeled training samples from unseen subjects and unseen scenarios video stimuli.

Method	**Metric**	**1-shot**	**2-shot**	**3-shot***	**4-shot***	**5-shot***
Eye movement	F1-score	0.12	0.14	0.14	0.14	0.15
	Accuracy	0.16	0.17	0.17	0.17	0.18
Periocular	F1-score	0.52	0.52	0.53	0.54	0.55
	Accuracy	0.52	0.52	0.53	0.54	0.55
Multi-faceted	F1-score	0.57	0.59	0.59	0.61	0.62
	Accuracy	0.58	0.60	0.60	0.61	0.62

*: For emotion class *Surprise*, at most two shots (i.e., two training instances) are taken.

distinguishing between neutral and surprise, as well as between high-arousal emotions such as disgust and fear. In the *cross-session settings*, performance generally degraded. Both methods exhibit consistent confusion between neutral and surprise, as well as among disgust, anger, and fear.

These confusions are partly due to the fact that these emotions often exhibit similar eye movement patterns and share overlapping visual cues in the periocular region. For instance, research has shown that expressions of anger and disgust share similar activation of facial muscles in the eye area [33], leading to potential misclassifications when relying solely on periocular features. Another reason for the observed confusion is that many of these emotions are not experienced in isolation. In practice, participants often feel a mixture of emotional responses. For example, the same video stimulus might simultaneously evoke both disgust and fear, or a combination of disgust and anger, as reported in previous studies [137, 138]. This is because these emotions are psychologically and physiologically interconnected. As a result, even when using precise Ekman's model, it is still difficult for the participant to assign a single label to present their emotional states, particularly for high-arousal or negatively valenced emotions such as anger, disgust, and fear.

5.10 Few-shot Performance in Cross-study Evaluation

To further investigate model's generalization capability in the context of unseen subjects and unseen emotional stimuli, we conduct a second few-shot learning evaluation. Specifically, each method is first pre-trained on data collected from Study A (20 subjects). We then fine-tune the pre-trained model using samples from Study B, which consists of six new subjects and a distinct set of emotional video stimuli. The evaluation is performed in a subject-specific manner: for each subject in Study B, we fine-tune the model using only a few-shot samples from that specific subject and test it on the remaining samples from the same subject. The final performance is reported as the average across all 6 subjects. The results are shown in Table 8.

As shown, the multi-faceted model consistently outperforms the eye movement and periocular baselines in all cases. Notably, even in the 1-shot setting, the multi-faceted model achieves an F1-score of 0.57 and an accuracy of 0.58, significantly surpassing the other methods. This performance advantage remains robust as the number of shots increases, reaching an F1-score of 0.62 and accuracy of 0.62 in the 5-shot setting. These results again suggest that the multi-faceted approach holds promise for improving generalization in real-world scenarios, especially when training data from the targeted user is scarce. However, there remains room for improvement in recognition performance. Future work could explore more advanced fusion strategies, adaptive fine-tuning mechanisms, or few-shot learning techniques to further enhance cross-subject and cross-context generalization.

6 DISCUSSION

6.1 Applications and Future Work

6.1.1 Multi-modal Sentiment Analysis in VR. Our benchmark study focuses on using diverse eye-tracking data, such as gaze, pupil diameter, and periocular images, for emotion recognition. In addition to this data, our dataset also includes high-frame-rate scene recordings of participants' view within the virtual environment when they engage with emotional stimuli, as well as the original audiovisual content of these stimuli. Together, this fusion of multi-modal information, supplemented by emotional labels and intensity ratings, provides a rich foundation for future multi-modal visual sentiment analysis in VR. Conventional visual sentiment analysis [119, 124, 132] often relies on visual cues from the stimuli to assess users' emotional responses. Our dataset, however, supports a more comprehensive multi-modal sentiment analysis [45, 119] by combining audiovisual stimuli with signals of attention (derived from gaze), excitement (from pupil diameter), and periocular micro-movements. This line of research allows for a more nuanced understanding of how VR content impacts user sentiment, which can further benefit applications such as personalized content adaptation, where VR experiences and content are dynamically tailored based on the user's real-time emotional responses and an accurate understanding of the VR content. Additionally, our dataset can support the development of multi-modal analytics systems that analyze emotional coherence across users in VR [126].

6.1.2 Benchmark for Privacy-preserving Emotion Recognition in VR. The potential leakage of sensitive information from raw eye images poses a major obstacle to allowing full access to eye trackers in modern head-mounted devices [22, 27]. A common solution to this problem is to prohibit any third-party application from accessing the raw data (with exceptions, such as devices with Tobii eye-tracking modules, where third-party applications can still obtain full access by purchasing the Tobii SDK license). However, as demonstrated in our benchmark study, certain applications, such as emotion recognition, rely on (and benefit significantly from) the raw periocular images to achieve effective functionality and maintain acceptable performance. Simply denying access to raw eye-tracking data would significantly limit the capabilities of legitimate third-party applications that require such information. Therefore, similar to ongoing efforts in privacy-preserving techniques for speech [25, 34] and face image-based emotion recognition [50, 79], it is essential to develop privacy-preserving solutions that eliminate sensitive user information from raw eye-tracking data while still enabling legitimate emotion recognition classifiers to operate effectively. However, there are currently no publicly available datasets that contain rich, raw eye-tracking data paired with emotion labels, creating a significant gap for advancing research in this area. Our dataset will bridge this research gap, and offer the research community a valuable resource for advancing privacy-compliant emotion tracking in VR and beyond. This includes developing anonymization or data abstraction techniques that retain essential features for emotion recognition while reducing privacy risks.

6.1.3 Advanced Emotion Recognition in VR. The preliminary results from our benchmark (Section 5) show that fusing multiple eye-tracking data types, i.e., gaze direction, pupil dynamics, and periocular images, can enhance emotion recognition, achieving an F1-score of 0.7 in both cross-session (unseen emotion stimuli) and few-shot learning scenarios. While promising, this level of accuracy highlights the need for further advancements, as practical VR applications demand higher, more reliable performance to ensure consistent accuracy across varying emotional contexts and user conditions. Our dataset, with its high-resolution, multi-faceted eye-tracking data, is designed to support this line of research by enabling new developments in deep learning and recognition algorithm design. For instance, the diverse signal types in the dataset provide opportunities to explore more effective multi-modal fusion methods, such as attention-based fusion [44] or cross-modality attention [62], to maximize the complementary nature of gaze, pupil, and periocular information. Moreover, our dataset encourages research into domain adaptation techniques tailored for emotion recognition, which can improve model robustness across different stimuli and user groups. By providing detailed and varied data, our dataset enables researchers to develop models that are not only more accurate but also adaptable for real-world VR applications where user and contextual diversity introduce practical challenges.

6.1.4 Emotion-aware Applications for Immersive Computing Systems. The dataset can help in developing emotion-aware applications for next-generation head-mounted immersive platforms (both augmented and virtual reality systems). Specifically, by leveraging the rich eye-tracking and periocular data in our dataset, emotion-aware systems can infer subtle affective cues in real time, enabling more natural, human-like interactions that adapt continuously to the user's emotional context. Such capabilities are especially valuable in training, education, and mental health support, where emotionally intelligent responses can improve user engagement, reduce cognitive load, and foster trust [87]. For example, we can develop dynamic game engines in VR that can adapt narratives, character behavior, or the progression of the storyline based on the real-time emotional state and intensity of the user as inputs. This capability will support the creation of deeply immersive, engaging, and personalized experiences in future interactive systems and gaming [46, 90]. Additionally, the dataset can support the development of emotionally adaptive virtual assistants [101, 117] that dynamically adjust their tone, pace, and interaction style in response to users' emotional states. For example, speaking more slowly and calmly when frustration is detected [67]. Furthermore, the dataset can enable personalized therapeutic experiences on immersive computing platforms [32, 57]. By leveraging real-time emotion recognition from eye movements and periocular cues, therapy sessions can be adapted to match the user's emotional state. In anxiety and phobia exposure therapy, emotion feedback can be used to adjust difficulty or pace in real time, ensuring user comfort and safety throughout the session.

6.1.5 Open-source Tool for Multi-modal Data Collection in VR. Alongside the dateset, we are also open-sourcing the data collection software tool (Section 3), which is designed to streamline future multi-modal data collection in VR. This comprehensive tool includes software components that other researchers can easily use to elicit emotional responses in virtual environments, reliably monitor and record diverse multi-modal signals from the VR device, and simplify the data annotation process. Specifically, key features of our software tool include a customizable emotional stimulus presentation module, allowing researchers to design and manage virtual scenarios that evoke specific emotional reactions. Additionally, the tool supports high-frequency signal processing and recording, ensuring that the multi-modal, high-resolution signals are captured accurately and synchronously for further analysis. A user-friendly data labeling interface is also included, enabling collaborative data annotation between participants and researchers, enhancing data accuracy and consistency.

By open-sourcing this tool, we aim to foster broader research initiatives in emotion recognition and multi-modal VR applications. For example, researchers can easily use our software tool to build a large-scale VR emotion dataset by incorporating the recent immersive and interactive 360-degree video dataset [51], to evoke emotional responses along arousal and valence dimensions.

6.2 Limitations

Our work can be further improved in the following ways.

Emotions beyond the Ekman's model. Our current work focuses on Ekman's seven basic emotions. While discrete emotional states enable direct mapping from the recognition results to downstream emotion-aware VR applications, i.e., using the detected emotion class of a VR user to generate expressive avatars with corresponding emotional expressions [3, 7, 82, 118, 128], it is important for future research to expand the emotion taxonomy and include more nuanced social or situational emotions, such as embarrassment, guilt, or pride. These complex emotions play a central role in real-world experiences and daily social interactions.

However, eliciting such emotions in a consistent and controlled manner remains an open challenge, as there is currently a lack of scientifically validated visual stimuli specifically designed to induce these emotions. Unlike the basic emotions, which are considered biologically universal and can be reliably evoked through audiovisual content, social and situational emotions are highly context-dependent and often involve an additional layer of cognitive appraisal or social reasoning [75]. Moreover, many of these complex emotions are not entirely independent from the basic emotions considered in this work; rather, they are often conceptualized as blends or contextual modulations of core affective states [100, 120]. For example, guilt and embarrassment may involve elements of sadness and fear, while pride may build upon happiness combined with surprise or self-awareness [120]. This hierarchical or compositional view of emotional experience supports our idea that robust recognition of the seven basic emotions can serve as a necessary foundation for modeling and interpreting more complex affective phenomena. As a direction for future work, we see strong potential in extending our current methodology by designing reliable, ethically approved stimuli that can elicit these more complex emotional states. Researchers can leverage the flexible data collection framework introduced in this work to explore a broader emotional spectrum in immersive environments and build richer, context-aware emotion recognition systems.

Age Limitations and Elderly Inclusion. A limitation of the current study lies in the age range of participants, which currently includes young adults from our university community. While this demographic is commonly used in early-stage VR and affective computing research [42, 43, 122], it limits the generalizability of our dataset to elderly populations. Prior research in affective science has shown that older adults often exhibit distinct emotional responses compared to younger individuals [47]. For instance, they tend to regulate emotions more effectively, focus more on positive stimuli, and demonstrate reduced physiological reactivity and altered gaze behavior, such as less fixation on negative emotional cues [48]. These age-related shifts in emotion perception and attention could significantly affect eye movement and periocular patterns. To address this limitation, future work can explore the use of suitable affective stimuli, enabling the safe and ethical inclusion of older participants.

Audio Data and Future Opportunities. As described in Section 3, our data collection system also captures audio recordings of participants. The raw audio recordings contain background noise, which primarily from spontaneous communication between participants and the researcher. We believe these audio signals still hold significant potential for future research, as they contain meaningful acoustic cues such as tone, pitch, and speech dynamics that can contribute to emotion recognition tasks. With appropriate noise filtering and audio segmentation, clean audio snippets can be extracted for use in multimodal analysis. We will include the audio data as part of the public dataset release to support further exploration of multimodal emotion recognition in immersive environments.

Emotion Elicitation Through Fully Immersive VR Stimuli. Although the emotional stimuli used in this study are drawn from a well-validated database and have been shown to elicit strong emotional responses [138], a potential limitation lies in the rendering of 2D video content within a 3D VR environment. Despite our design efforts to optimize visual immersion through curved display and field-of-view alignment, the experience may not fully replicate the spatial immersion provided by native 360° or interactive VR content. Future work can leverage the open-source framework developed in this work to explore the integration of fully immersive VR-native stimuli to further enhance emotional presence.

7 CONCLUSION

We presented a comprehensive eye-tracking dataset specifically designed for emotion recognition research in VR, along with a data collection software tool and extensive benchmarking analysis. Our dataset is the first to provide periocular eye videos and Ekman's basic emotion annotations, distinguishing it from previous VR-based emotion recognition datasets. As such, it addresses key research gaps and enables new research in emotion recognition and affect-aware computing in VR. To validate its effectiveness, we conducted a benchmark evaluation demonstrating that fusing the new periocular video data available in our dataset with conventional eye-tracking signals, such as gaze direction and pupil dynamics, significantly enhances emotion recognition performance. Our data collection software, relying solely on standard sensors found in most AR headsets, ensures compatibility and scalability across commercial platforms. We believe this work provides a valuable, scalable resource for exploring emotional and cognitive processes in VR, paving the way for more emotionally responsive and immersive VR experiences.

8 DATA AND CODE AVAILABILITY

We are open-sourcing all research artifacts from this work, including the data collection software framework, video stimuli, and scripts for data processing and deep learning models used in benchmarking, to enhance reusability and reproducibility. These resources are available at https://github.com/MultiRepEyeVR/Through-the-Eyes-of-Emotion. The dataset is maintained in our institutional repository and is available upon request following the signing of a sharing agreement.

Acknowledgments

We would like to express our gratitude to the anonymous reviewers and the associate editors for their insightful comments and guidance. This work was supported in part by the Meta Research Award, SURF Research Cloud grant EINF-6360, and the EU's Horizon Europe HarmonicAI project under the HORIZON MSCA-2022-SE-01 scheme with grant agreement number 101131117. The contents of this paper do not necessarily reflect the positions or policies of the funding agencies.

References

[1] Mariano Alcañiz, Rosa Baños, Cristina Botella, and Beatriz Rey. 2003. The EMMA Project: Emotions as a determinant of presence. *PsychNology Journal* 1, 2 (2003), 141–150.

[2] Devon Allcoat and Adrian von Mühlenen. 2018. Learning in virtual reality: Effects on performance, emotion and engagement. *Research in Learning Technology* 26 (2018).

[3] Leonardo Angelini, Massimo Mecella, Hai-Ning Liang, Maurizio Caon, Elena Mugellini, Omar Abou Khaled, and Danilo Bernardini. 2022. Towards an emotionally augmented metaverse: A framework for recording and analysing physiological data and user behaviour. In *Proceedings of the 13th Augmented Human International Conference*. 1–5.

[4] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6836–6846.

[5] Swarnali Banik, Sougata Sen, Snehanshu Saha, and Surjya Ghosh. 2024. Towards reducing continuous emotion annotation effort during video consumption: A physiological response profiling approach. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–32.

[6] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the ACM International Conference on Multimodal Interaction*.

[7] Guillermo Bernal and Pattie Maes. 2017. Emotional beasts: Visually expressing emotions through avatars in VR. In *Proceedings of the CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 2395–2402.

[8] David Bethge, Luis Falconeri Coelho, Thomas Kosch, Satiyabooshan Murugaboopathy, Ulrich von Zadow, Albrecht Schmidt, and Tobias Grosse-Puppendahl. 2023. Technical design space analysis for unobtrusive driver emotion assessment using multi-domain context. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–30.

[9] Margaret M Bradley, Maurizio Codispoti, Dean Sabatinelli, and Peter J Lang. 2001. Emotion and motivation II: Sex differences in picture processing. *Emotion* 1, 3 (2001), 300.

[10] 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 (1994), 49–59.

[11] Margaret M Bradley, Laura Miccoli, Miguel A Escrig, and Peter J Lang. 2008. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology* 45, 4 (2008), 602–607.

[12] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, Benjamin Weiss, et al. 2005. A database of german emotional speech. In *Proceedings of the Interspeech Conference*. 1517–1520.

[13] Rafael A Calvo, Sidney D'Mello, Jonathan Matthew Gratch, and Arvid Kappas. 2015. *The Oxford handbook of affective computing*. Oxford University Press, USA.

[14] David Checa and Andres Bustillo. 2020. A review of immersive virtual reality serious games to enhance learning and training. *Multimedia Tools and Applications* 79, 9 (2020), 5501–5527.

[15] Hao Chen, Arindam Dey, Mark Billinghurst, and Robert W Lindeman. 2017. Exploring the design space for multi-sensory heart rate feedback in immersive virtual reality. In *Proceedings of the Australian Conference on Computer-Human Interaction*. 108–116.

[16] Yih-Giun Cherng, Talia Baird, Jui-Tai Chen, and Chin-An Wang. 2020. Background luminance effects on pupil size associated with emotion and saccade preparation. *Scientific Reports* 10, 1 (2020), 15718.

[17] Luca Chittaro and Fabio Buttussi. 2015. Assessing knowledge retention of an immersive serious game vs. a traditional education method in aviation safety. *IEEE Transactions on Visualization and Computer Graphics* 21, 4 (2015), 529–538.

[18] Gonçalo Cosme, Pedro J Rosa, César F Lima, Vânia Tavares, Sophie Scott, Sinead Chen, Thomas DW Wilcockson, Trevor J Crawford, and Diana Prata. 2021. Pupil dilation reflects the authenticity of received nonverbal vocalizations. *Scientific Reports* 11, 1 (2021), 3733.

[19] Sandra Costa, Alberto Brunete, Byung-Chull Bae, and Nikolaos Mavridis. 2018. Emotional storytelling using virtual and robotic agents. *International Journal of Humanoid Robotics* 15, 03 (2018), 1850006.

[20] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou, Edelle McMahon, Martin Sawey, and Marc Schröder. 2000. 'FEELTRACE': An instrument for recording perceived emotion in real time. In *Proc. ITRW on Speech and Emotion*. 19–24.

[21] Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard. 2021. Attentional feature fusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3560–3569.

[22] Brendan David-John, Diane Hosfelt, Kevin Butler, and Eakta Jain. 2021. A privacy-preserving approach to streaming eye-tracking data. *IEEE Transactions on Visualization and Computer Graphics* 27, 5 (2021), 2555–2565.

[23] Nele AJ De Witte, Sara Scheveneels, Romy Sels, Glen Debard, Dirk Hermans, and Tom Van Daele. 2020. Augmenting exposure therapy: Mobile augmented reality for specific phobia. *Frontiers in Virtual Reality* 1 (2020), 8.

[24] Arindam Dey, Thammathip Piumsomboon, Youngho Lee, and Mark Billinghurst. 2017. Effects of sharing physiological states of players in a collaborative virtual reality gameplay. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*. 4045–4056.

[25] Miguel Dias, Alberto Abad, and Isabel Trancoso. 2018. Exploring hashing and cryptonet based approaches for privacy-preserving speech emotion recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. 2057–2061.

[26] Miranda R Donnelly, Renee Reinberg, Kaori L Ito, David Saldana, Meghan Neureither, Allie Schmiesing, Esther Jahng, and Sook-Lei Liew. 2021. Virtual reality for the treatment of anxiety disorders: A scoping review. *The American Journal of Occupational Therapy* 75, 6 (2021).

[27] Lingyu Du, Jinyuan Jia, Xucong Zhang, and Guohao Lan. 2024. PrivateGaze: Preserving User Privacy in Black-box Mobile Gaze Tracking Services. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–28.

[28] Paul Ekman. 1994. Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychological Bulletin* 115, 2 (1994), 268–287.

[29] Paul Ekman and Wallace Friesen. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology* 17 (02 1971), 124–9.

[30] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition* 44, 3 (2011), 572–587.

[31] Paul MG Emmelkamp and Katharina Meyerbröker. 2021. Virtual reality therapy in mental health. *Annual review of clinical psychology* 17, 1 (2021), 495–519.

[32] Iveta Fajnerova, Lukáš Hejtmánek, Michal Sedlák, Markéta Jablonská, Anna Francová, and Pavla Stopková. 2024. The journey from nonimmersive to immersive multiuser applications in mental health care: Systematic review. *Journal of Medical Internet Research* 26 (2024), e60441.

[33] Larissa L Faustmann, Lara Eckhardt, Pauline S Hamann, and Mareike Altgassen. 2022. The effects of separate facial areas on emotion recognition in different adult age groups: A laboratory and a naturalistic study. *Frontiers in Psychology* 13 (2022), 859464.

[34] Tiantian Feng, Hanieh Hashemi, Murali Annavaram, and Shrikanth S Narayanan. 2022. Enhancing privacy through domain adaptive noise injection for speech emotion recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. 7702–7706.

[35] Crystal A Gabert-Quillen, Ellen E Bartolini, Benjamin T Abravanel, and Charles A Sanislow. 2015. Ratings for emotion film clips. *Behavior Research Methods* 47 (2015), 773–787.

[36] Christoph Gebhardt, Andreas Brombach, Tiffany Luong, Otmar Hilliges, and Christian Holz. 2024. Detecting users' emotional states during passive social media use. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (2024), 1–30.

[37] Adélaïde Genay, Anatole Lécuyer, and Martin Hachet. 2021. Being an avatar "for real": A survey on virtual embodiment in augmented reality. *IEEE Transactions on Visualization and Computer Graphics* 28, 12 (2021), 5071–5090.

[38] T Lee Gilman, Razan Shaheen, K Maria Nylocks, Danielle Halachoff, Jessica Chapman, Jessica J Flynn, Lindsey M Matt, and Karin G Coifman. 2017. A film set for the elicitation of emotion in research: A comprehensive catalog derived from four decades of investigation. *Behavior Research Methods* 49 (2017), 2061–2082.

[39] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

[40] Lynsey Gregg and Nicholas Tarrier. 2007. Virtual reality in mental health: A review of the literature. *Social Psychiatry and Psychiatric Epidemiology* 42 (2007), 343–354.

[41] James J Gross and Robert W Levenson. 1995. Emotion elicitation using films. *Cognition & Emotion* 9, 1 (1995), 87–108.

[42] Quentin Guimard, Florent Robert, Camille Bauce, Aldric Ducreux, Lucile Sassatelli, Hui-Yin Wu, Marco Winckler, and Auriane Gros. 2022. PEM360: A dataset of 360° videos with continuous physiological measurements, subjective emotional ratings and motion traces. In *Proceedings of the ACM Multimedia Systems Conference*. 252–258.

[43] 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 (2022), 1–21.

[44] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. 2017. Attention-based multimodal fusion for video description. In *Proceedings of the IEEE International Conference on Computer Vision*. 4193–4202.

[45] Anthony Hu and Seth Flaxman. 2018. Multimodal sentiment analysis to explore the structure of emotions. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 350–358.

[46] Eva Hudlicka. 2009. Affective game engines: Motivation and requirements. In *Proceedings of the 4th International Conference on Foundations of Digital Games*. 299–306.

[47] Derek M Isaacowitz, Kimberly M Livingstone, and Vanessa L Castro. 2017. Aging and emotions: Experience, regulation, and perception. *Current Opinion in Psychology* 17 (2017), 79–83.

[48] Derek M Isaacowitz, Heather A Wadlinger, Deborah Goren, and Hugh R Wilson. 2006. Selective preference in visual fixation away from negative images in old age? An eye-tracking study. *Psychology and Aging* 21, 1 (2006), 40.

[49] Carroll E. Izard. 1977. *Human Emotions*. New York: Springer US.

[50] Mimansa Jaiswal and Emily Mower Provost. 2020. Privacy enhanced multimodal neural representations for emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 7985–7993.

[51] 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* (2024).

[52] Stamos Katsigiannis and Naeem Ramzan. 2017. DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE Journal of Biomedical and Health Informatics* 22, 1 (2017), 98–107.

[53] Panayu Keelawat, Nattapong Thammasan, Masayuki Numao, and Boonserm Kijsirikul. 2021. A comparative study of window size and channel arrangement on EEG-emotion recognition using deep CNN. *Sensors* 21, 5 (2021), 1678.

[54] Valerie L Kinner, Lars Kuchinke, Angelika M Dierolf, Christian J Merz, Tobias Otto, and Oliver T Wolf. 2017. What our eyes tell us about feelings: Tracking pupillary responses during emotion regulation processes. *Psychophysiology* 54, 4 (2017), 508–518.

[55] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. DEAP: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing* 3, 1 (2011), 18–31.

[56] Mariska E Kret and Elio E Sjak-Shie. 2019. Preprocessing pupil size data: Guidelines and code. *Behavior Research Methods* 51 (2019), 1336–1342.

[57] Jacob Kritikos, Georgios Alevizopoulos, and Dimitris Koutsouris. 2021. Personalized virtual reality human-computer interaction for psychiatric and neurological illnesses: A dynamically adaptive virtual reality environment that changes according to real-time feedback from electrophysiological signal responses. *Frontiers in Human Neuroscience* 15 (2021), 596980.

[58] Pin-Sung Ku, Yu-Chih Lin, Yi-Hao Peng, and Mike Chen. 2019. PeriText: Utilizing Peripheral Vision for Reading Text on Augmented Reality Smart Glasses. 630–635.

[59] Gyouhyung Kyung and Sungryul Park. 2021. Curved Versus Flat Monitors: Interactive Effects of Display Curvature Radius and Display Size on Visual Search Performance and Visual Fatigue. *Human Factors* 63, 7 (2021), 1182–1195.

[60] Guohao Lan, Bailey Heit, Tim Scargill, and Maria Gorlatova. 2020. GazeGraph: Graph-based few-shot cognitive context sensing from human visual behavior. In *Proceedings of the ACM Conference on Embedded Networked Sensor Systems*. 422–435.

[61] Guohao Lan, Tim Scargill, and Maria Gorlatova. 2022. EyeSyn: Psychology-inspired eye movement synthesis for gaze-based activity recognition. In *Proceedings of ACM/IEEE International Conference on Information Processing in Sensor Networks*. IEEE, 233–246.

[62] Wei-Yu Lee, Ljubomir Jovanov, and Wilfried Philips. 2022. Cross-modality attention and multimodal fusion transformer for pedestrian detection. In *Proceedings of European Conference on Computer Vision*. Springer, 608–623.

[63] 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 (2017), 2116.

[64] 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 (2022), 3832–3842.

[65] Yi Li, Adel S Elmaghraby, Ayman El-Baz, and Estate M Sokhadze. 2015. Using physiological signal analysis to design affective VR games. In *Proceedings of IEEE International Symposium on Signal Processing and Information Technology*. 57–62.

[66] Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. 2022. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *Proceedings of the ACM International Conference on Multimedia*. 24–32.

[67] 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 (2022), 3917–3927.

[68] Tiffany Luong, Anatole Lecuyer, Nicolas Martin, and Ferran Argelaguet. 2021. A survey on affective and cognitive VR. *IEEE Transactions on Visualization and Computer Graphics* 28, 12 (2021), 5154–5171.

[69] Mary F Macedonio, Thomas D Parsons, Raymond A Digiuseppe, Brenda A Weiderhold, and Albert A Rizzo. 2007. Immersiveness and physiological arousal within panoramic video-based virtual reality. *Cyberpsychology & Behavior* 10, 4 (2007), 508–515.

[70] Javier Marín-Morales, Juan Luis Higuera Trujillo, Alberto Greco, Jaime Guixeres, Carmen Llinares, Enzo Scilingo, Mariano Alcañiz Raya, and Gaetano Valenza. 2018. Affective computing in virtual reality: Emotion recognition from brain and heartbeat dynamics using wearable sensors. *Scientific Reports* 8 (09 2018).

[71] 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 (2020).

[72] Sebastiaan Mathôt, Jasper Fabius, Elle Van Heusden, and Stefan Van der Stigchel. 2018. Safe and sensible preprocessing and baseline correction of pupil-size data. *Behavior Research Methods* 50 (2018), 94–106.

[73] David Matsumoto and Hyi Sung Hwang. 2012. Culture and emotion: The integration of biological and cultural contributions. *Journal of Cross-Cultural Psychology* 43, 1 (2012), 91–118.

[74] Magdalena Matyjek, Mareike Bayer, and Isabel Dziobek. 2021. Pupillary responses to faces are modulated by familiarity and rewarding context. *Brain Sciences* 11, 6 (2021), 794.

[75] Kateri McRae, S Megan Heller, Oliver P John, and James J Gross. 2011. Context-dependent emotion regulation: Suppression and reappraisal at the Burning Man festival. *Basic and Applied Social Psychology* 33, 4 (2011), 346–350.

[76] Albert Mehrabian and James A. Russell. 1974. The basic emotional impact of environments. *Perceptual and Motor Skills* 38, 1 (1974), 283–301.

[77] Mohammadhossein Moghimi, Robert Stone, and Pia Rotshtein. 2017. Affective recognition in dynamic and interactive virtual environments. *IEEE Transactions on Affective Computing* 11, 1 (2017), 45–62.

[78] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* 10, 1 (2017), 18–31.

[79] Vansh Narula, Kexin Feng, and Theodora Chaspari. 2020. Preserving privacy in image-based emotion recognition through user anonymization. In *Proceedings of the International Conference on Multimodal Interaction*. 452–460.

[80] Manuel Oliva and Andrey Anikin. 2018. Pupil dilation reflects the time course of emotion recognition in human vocalizations. *Scientific Reports* 8, 1 (2018), 4871.

[81] SREEJA P S and Mahalakshmi G S. 2017. Emotion Models: A Review. *International Journal of Control Theory and Applications* 10 (01 2017), 651–657.

[82] Ye Pan, Shuai Tan, Shengran Cheng, Qunfen Lin, Zijiao Zeng, and Kenny Mitchell. 2024. Expressive talking avatars. *IEEE Transactions on Visualization and Computer Graphics* (2024).

[83] Eunji Park, Duri Lee, Yunjo Han, James Diefendorff, and Uichin Lee. 2024. Hide-and-seek: Detecting workers' emotional workload in emotional labor contexts using multimodal sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–28.

[84] 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 (2023), 238–245.

[85] Farrukh Pervez, Moazzam Shoukat, Muhammad Usama, Moid Sandhu, Siddique Latif, and Junaid Qadir. 2024. Affective Computing and the Road to an Emotionally Intelligent Metaverse. *IEEE Open Journal of the Computer Society* (2024).

[86] Pierre Philippot. 1993. Inducing and assessing differentiated emotion-feeling states in the laboratory. *Cognition and Emotion* 7, 2 (1993), 171–193.

[87] Rosalind W Picard. 2008. Toward machines with emotional intelligence. (2008).

[88] R Gnana Praveen, Wheidima Carneiro de Melo, Nasib Ullah, Haseeb Aslam, Osama Zeeshan, Théo Denorme, Marco Pedersoli, Alessandro L Koerich, Simon Bacon, Patrick Cardinal, et al. 2022. A joint cross-attention model for audio-visual fusion in dimensional emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2486–2495.

[89] Gabriela Maria Pyjas, Jonathan Weinel, and Martyn Broadhead. 2022. Storytelling and VR: Inducing emotions through AI characters. In *Proceedings of EVA London*. BCS Learning & Development, 198–204.

[90] Niklas Ravaja, Timo Saari, Mikko Salminen, Jari Laarni, and Kari Kallinen. 2006. Phasic emotional reactions to video game events: A psychophysiological investigation. *Media Psychology* 8, 4 (2006), 343–367.

[91] Narsi Reddy and Reza Derakhshani. 2020. Emotion detection using periocular region: A cross-dataset study. In *Proceedings of the International Joint Conference on Neural Networks*. IEEE, 1–6.

[92] Daniel Roth, Jean-Luc Lugrin, Dmitri Galakhov, Arvid Hofmann, Gary Bente, Marc Erich Latoschik, and Arnulph Fuhrmann. 2016. Avatar realism and social interaction quality in virtual reality. In *Proceedings of IEEE Virtual Reality*. IEEE, 277–278.

[93] James A Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161.

[94] James A Russell. 1994. Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin* 115, 1 (1994), 102.

[95] Heini Saarimäki. 2021. Naturalistic stimuli in affective neuroimaging: A review. *Frontiers in Human Neuroscience* 15 (2021), 675068.

[96] Asmaa Sakr and Tariq Abdullah. 2024. Virtual, augmented reality and learning analytics impact on learners, and educators: A systematic review. *Education and Information Technologies* (2024), 1–50.

[97] Atanu Samanta and Tanaya Guha. 2020. Emotion sensing from head motion capture. *IEEE Sensors Journal* 21, 4 (2020), 5035–5043.

[98] Andrea C Samson, Sylvia D Kreibig, Blake Soderstrom, A Ayanna Wade, and James J Gross. 2016. Eliciting positive, negative and mixed emotional states: A film library for affective scientists. *Cognition and Emotion* 30, 5 (2016), 827–856.

[99] Alexandre Schaefer, Frédéric Nils, Xavier Sanchez, and Pierre Philippot. 2010. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and Emotion* 24, 7 (2010), 1153–1172.

[100] Klaus R Scherer. 2005. What are emotions? And how can they be measured? *Social Science Information* 44, 4 (2005), 695–729.

[101] Andreas Schmeil and Wolfgang Broll. 2007. Mara-a mobile augmented reality-based virtual assistant. In *Proceedings of IEEE Virtual Reality Conference*. IEEE, 267–270.

[102] Caspar M Schwiedrzik and Sandrin S Sudmann. 2020. Pupil diameter tracks statistical structure in the environment to increase visual sensitivity. *Journal of Neuroscience* 40, 23 (2020), 4565–4575.

[103] Henrique Siqueira, Sven Magg, and Stefan Wermter. 2020. Efficient facial feature learning with wide ensemble-based convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 5800–5809.

[104] Vasileios Skaramagkas, Giorgos Giannakakis, Emmanouil Ktistakis, Dimitris Manousos, Ioannis Karatzanis, Nikolaos S Tachos, Evanthia Tripoliti, Kostas Marias, Dimitrios I Fotiadis, and Manolis Tsiknakis. 2021. Review of eye tracking metrics involved in emotional and cognitive processes. *IEEE Reviews in Biomedical Engineering* 16 (2021), 260–277.

[105] Jamie Snytte, Ting Ting Liu, Renée Withnell, M Natasha Rajah, and Signy Sheldon. 2025. Emotional events induce retrograde memory impairments on conceptually-related neutral events. *Cognition* 259 (2025), 106103.

[106] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. 2011. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing* 3, 1 (2011), 42–55.

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

[108] Joep Sonnemans and Nico H Frijda. 1994. The structure of subjective emotional intensity. *Cognition & Emotion* 8, 4 (1994), 329–350.

[109] Stuart R Steinhauer, Margaret M Bradley, Greg J Siegle, Kathryn A Roecklein, and Annika Dix. 2022. Publication guidelines and recommendations for pupillary measurement in psychophysiological studies. *Psychophysiology* 59, 4 (2022), e14035.

[110] Ekaterina Sviridova, Elena Yastrebova, Gulmira Bakirova, and Fayruza Rebrina. 2023. Immersive technologies as an innovative tool to increase academic success and motivation in higher education. *Frontiers in Education* 8 (2023), 1192760.

[111] Luma Tabbaa, Ryan Searle, Saber Mirzaee Bafti, Md Moinul Hossain, Jittrapol Intarasisrisawat, Maxine Glancy, and Chee Siang Ang. 2021. VREED: Virtual reality emotion recognition dataset using eye tracking & physiological measures. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–20.

[112] Ying Tan, Zhe Sun, Feng Duan, Jordi Solé-Casals, and Cesar F Caiafa. 2021. A multimodal emotion recognition method based on facial expressions and electroencephalography. *Biomedical Signal Processing and Control* 70 (2021), 103029.

[113] Paweł Tarnowski, Marcin Kołodziej, Andrzej Majkowski, Remigiusz Jan Rak, and Amparo Alonso-Betanzos. 2020. Eye tracking analysis for emotion recognition. 2020 (2020), 13 pages.

[114] Meike K Uhrig, Nadine Trautmann, Ulf Baumgärtner, Rolf-Detlef Treede, Florian Henrich, Wolfgang Hiller, and Susanne Marschall. 2016. Emotion elicitation: A comparison of pictures and films. *Frontiers in Psychology* 7 (2016), 180.

[115] Avinash R Vaidya, Chenshuo Jin, and Lesley K Fellows. 2014. Eyespy: The predictive value of fixation patterns in detecting subtle and extreme emotions from faces. *Cognition* 133, 2 (2014), 443–456.

[116] Rafael Wampfler, Severin Klingler, Barbara Solenthaler, Victor R Schinazi, Markus Gross, and Christian Holz. 2022. Affective state prediction from smartphone touch and sensor data in the wild. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*. 1–14.

[117] Isaac Wang, Jesse Smith, and Jaime Ruiz. 2019. Exploring virtual agents for augmented reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[118] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. 2020. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *Proceedings of European Conference on Computer Vision*. Springer, 700–717.

[119] Xingbo Wang, Jianben He, Zhihua Jin, Muqiao Yang, Yong Wang, and Huamin Qu. 2021. M2lens: Visualizing and explaining multimodal models for sentiment analysis. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 802–812.

[120] David Watson and Kasey Stanton. 2017. Emotion blends and mixed emotions in the hierarchical structure of affect. *Emotion Review* 9, 2 (2017), 99–104.

[121] Suowei Wu, Zhengyin Du, Weixin Li, Di Huang, and Yunhong Wang. 2019. Continuous emotion recognition in videos by fusing facial expression, head pose and eye gaze. In *Proceedings of the ACM International Conference on Multimodal Interactio*. 40–48.

[122] Tong Xue, Abdallah El Ali, Tianyi Zhang, Gangyi Ding, and Pablo Cesar. 2021. CEAP-360VR: A Continuous Physiological and Behavioral Emotion Annotation Dataset for 360VR Videos. *IEEE Transactions on Multimedia* 25 (2021), 243–255.

[123] Tong Xue, Abdallah El Ali, Tianyi Zhang, Gangyi Ding, and Pablo Cesar. 2021. RCEA-360VR: Real-time, continuous emotion annotation in 360° VR videos for collecting precise viewport-dependent ground truth labels. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*. 1–15.

[124] Jufeng Yang, Dongyu She, Yu-Kun Lai, Paul L Rosin, and Ming-Hsuan Yang. 2018. Weakly supervised coupled networks for visual sentiment analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7584–7592.

[125] Kangning Yang, Benjamin Tag, Chaofan Wang, Yue Gu, Zhanna Sarsenbayeva, Tilman Dingler, Greg Wadley, and Jorge Goncalves. 2022. Survey on emotion sensing using mobile devices. *IEEE Transactions on Affective Computing* 14, 4 (2022), 2678–2696.

[126] Haipeng Zeng, Xingbo Wang, Aoyu Wu, Yong Wang, Quan Li, Alex Endert, and Huamin Qu. 2019. EmoCo: Visual analysis of emotion coherence in presentation videos. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2019), 927–937.

[127] George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. 2021. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2114–2124.

[128] Bingyuan Zhang, Xulong Zhang, Ning Cheng, Jun Yu, Jing Xiao, and Jianzong Wang. 2024. Emotalker: Emotionally editable talking face generation via diffusion model. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. 8276–8280.

[129] 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 Conference Proceedings*. 1–11.

[130] Jing Zhang, Sung Park, Ayoung Cho, and Mincheol Whang. 2022. Significant Measures of Gaze and Pupil Movement for Evaluating Empathy between Viewers and Digital Content. *Sensors* 22, 5 (2022).

[131] Jianfeng Zhao, Xia Mao, and Lijiang Chen. 2019. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control* 47 (2019), 312–323.

[132] Yingying Zhao, Yuhu Chang, Yutian Lu, Yujiang Wang, Mingzhi Dong, Qin Lv, Robert P Dick, Fan Yang, Tun Lu, Ning Gu, et al. 2022. Do smart glasses dream of sentimental visions? Deep emotionship analysis for eyewear devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–29.

[133] Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. 2018. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics* 49, 3 (2018), 1110–1122.

[134] Peixiang Zhong, Di Wang, and Chunyan Miao. 2020. EEG-based emotion recognition using regularized graph neural networks. *IEEE Transactions on Affective Computing* 13, 3 (2020), 1290–1301.

[135] Zhiwei Zhu, Kikuo Fujimura, and Qiang Ji. 2002. Real-time eye detection and tracking under various light conditions. In *Proceedings of the Symposium on Eye tracking Research & Applications*. 139–144.

[136] Dolf Zillmann. 2008. Excitation transfer theory. *The International Encyclopedia of Communication* (2008).

[137] Barbra Zupan and Duncan R Babbage. 2017. Film clips and narrative text as subjective emotion elicitation techniques. *The Journal of Social Psychology* 157, 2 (2017), 194–210.

[138] Barbra Zupan and Michelle Eskritt. 2020. Eliciting emotion ratings for a set of film clips: A preliminary archive for research in emotion. *The Journal of Social Psychology* 160, 6 (2020), 768–789.

A Appendix: Information of Video Stimuli

This section describes the video stimuli used in our data collection process (introduced in Section 4). The dataset was collected in two stages of user studies. The first stage involved 20 participants, and the second stage involved six participants. In each stage, participants were shown a curated set of 14 short video clips designed to evoke the seven basic emotions. The video stimuli were selected based on the emotional intensity ratings reported in the prior study by Zupan et al. [138]. Emotional intensity was rated on a scale from 1 (not at all) to 9 (extremely).

For Study One, we selected clips with the highest intensity ratings for each target emotion, prioritizing both emotional impact and narrative clarity. Each video clip was accompanied by a brief content sentence presented to participants before playback to provide context and aid in emotional engagement. Study Two used a different set of 14 clips, also selected from the same source, with comparable emotional intensity values but different scenarios and character portrayals. This allowed us to evaluate the generalizability and robustness of participants' emotional responses across diverse stimuli. The complete list of video clips, along with their emotional intensity ratings, durations, and briefing content, is provided in Tables 9 and 10 for Study One and Study Two, respectively.

Table 9. Details of the 14 video clips used in Study One to evoke the seven basic emotions. Emotional intensity ranges from 1 (not at all) to 9 (extremely). The content sentence is used to brief the participants about the video clip before playing.

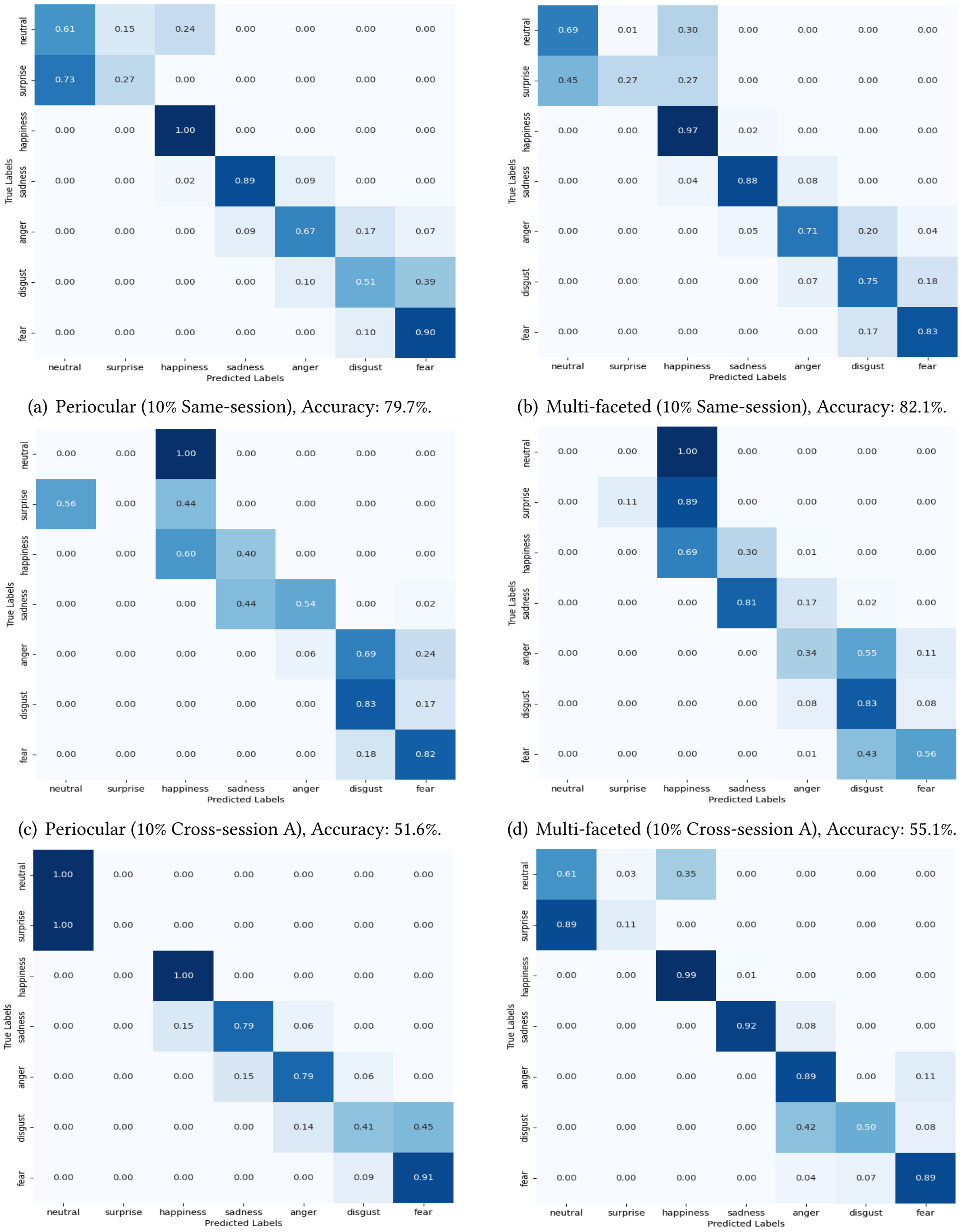
Emotion	Video clip	Emotional intensity	Length (s)	Content sentence
Neutral	(a) Ex Machina	6.8 (2.5)	43	A man was selected to assist another man in a research project. The two are discussing topics related to the job.
	(b) Rudderless (business meeting)	6.9 (2.6)	30	A group of coworkers are having a business meeting to try and secure a partnership with a fellow company.
Surprise	(a) One Day	8.2 (1.7)	38	A woman is riding her bike to meet her husband for a date.
	(b) Deep Blue Sea	7.7 (2.0)	94	A man tries to convince his fellow scientists to not give up when they find themselves trapped in the middle of the ocean.
Happiness	(a) Soul Surfer (homeless girl)	7.6 (1.9)	151	After a girl loses her arm in a shark attack while surfing, she goes on a mission trip with her church to help Tsunami victims.
	(b) Lottery Ticket	7.5 (1.9)	112	A young man was buying a lottery ticket for his grandmother when the store clerk convinced him to buy one for himself too.
Sadness	(a) Still Alice	7.5 (1.6)	142	A woman and her husband are spending time together at their beach house after they find out she has early onset Alzheimer's.
	(b) My Sister's Keeper	7.7 (1.7)	106	A girl is dying from terminal cancer and she doesn't want to fight anymore.
Anger	(a) 12 Years a Slave	7.8 (1.7)	115	A mother and her two children have been kidnapped by slave traders and are up for sale at a slave auction.
	(b) Enough	8.2 (1.5)	119	A young woman working in a diner for a meager wage marries her dream man and has a child.
Disgust	(a) American History X	8.3 (1.2)	149	After some men try to rob his car, a white supremacist takes action despite his young brother's protests.
	(b) Limitless (blood)	7.7 (2.0)	46	A gang has broken into a man's house to steal his drugs which unlock the full potential of one's brain. The man has run out of the drug and is suffering the withdrawals on the floor of his apartment while the gang searches for the tablets.
Fear	(a) Kings of Summer	6.7 (2.4)	61	Three friends ran away from home and built a cabin in the woods. Eventually, two of them return home but the third insists on staying and living alone.
	(b) The Conjuring	7.3 (2.4)	141	A woman hears something and heads into the basement to investigate.

Table 10. Details of the 14 video clips used in Study Two to evoke the seven basic emotions. The emotional intensity values (mean and standard deviation) are drawn from the study by Zupan et al. [138].

Emotion	Video clip	Emotional intensity	Length (s)	Content sentence
Neutral	(a) Good Will Hunting	5.5 (2.8)	88	A man and a woman go on a date at a greyhound racing track, where she asks about his family.
	(b) The Other Woman	5.4 (2.8)	45	A husband and wife discuss everyday matters: she updates him on household issues, and he talks about work and signing new contracts.
Surprise	(a) The Call (parking garage)	6.4 (2.4)	26	A young girl, walking through a parking garage, makes a phone call to her mother.
	(b) Joe	7.4 (2.1)	32	A group of butcher friends chat about an elderly woman's birthday. One man plans a trip to town to pick up items that others need.
Happiness	(a) Forrest Gump (reunion)	7.5 (1.7)	51	In the middle of a presentation, a man rushes off stage upon spotting the love of his life running toward him. They embrace in a pond near the Lincoln Memorial.
	(b) Soul Surfer (surfing)	7.4 (1.7)	109	A disabled girl who loves surfing heads to the beach with friends. After several attempts, she finally succeeds in catching a wave.
Sadness	(a) I Am Sam	7.9 (1.5)	108	A mentally challenged father is struggling in court as he fights to keep custody of his daughter.
	(b) My Sister's Keeper (doctor)	7.9 (1.5)	100	A young girl with leukemia hears a grim update from her doctor, who tells her mother that she may not have much time left.
Anger	(a) Crash	8.1 (1.1)	92	A black couple is pulled over by a racist cop. He treats the man violently and sexually harasses the woman.
	(b) The Hunting Ground	7.8 (1.7)	122	Several women recount their experiences of being raped and describe the frustration they faced when seeking help.
Disgust	(a) Wild	7.6 (2.0)	62	A woman takes an arduous hike until her feet become severely swollen, causing immense discomfort.
	(b) Slumdog Millionaire (blinded)	8.3 (1.4)	73	Two human traffickers force a boy to witness another child being drugged and then blinded with scalding liquid.
Fear	(a) The Life Before Her Eyes	7.2 (2.2)	120	Two girls apply makeup in a school bathroom when they hear screams from outside, followed by gunshots.
	(b) Insidious	6.95 (2.6)	151	A woman describes a chilling dream involving a child's room and a frightening presence, which suddenly appears and sends her running in terror.

B Appendix: Confusion Matrix

Figures 7 and 8 present the confusion matrices correspond to the recognition results reported in Table 7. As the evaluation follows a leave-one-subject-out protocol, the results are subject-dependent. Instead of showing the results for all the subjects, we take Subjects 3 and 7 as two representative examples. The figures below present the confusions between different emotional classes when the two models (periocular and multi-faceted) are fine-tuned by data with empirically estimated labels (10% data from the same-session or 10% data from cross-session).



(a) Periocular (10% Same-session), Accuracy: 79.7%.

(b) Multi-faceted (10% Same-session), Accuracy: 82.1%.

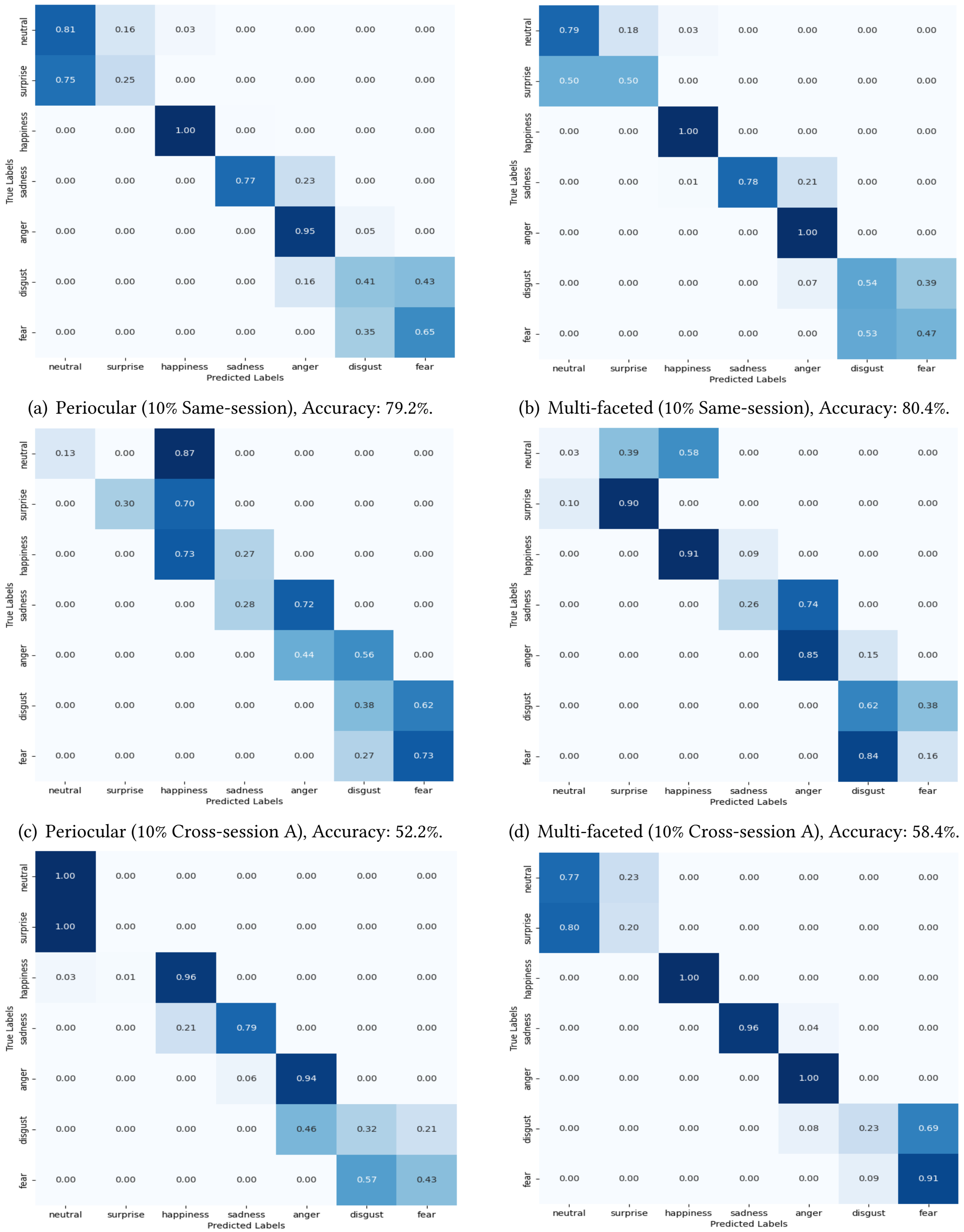
(c) Periocular (10% Cross-session A), Accuracy: 51.6%.

(d) Multi-faceted (10% Cross-session A), Accuracy: 55.1%.

(e) Periocular (10% Cross-session B), Accuracy: 78.2%.

(f) Multi-faceted (10% Cross-session B), Accuracy: 87.1%.

Fig. 7. Confusion matrices of the recognition results for Subject 3 when fine-tuning by data with empirically estimated labels. These results correspond to the setting and performance reported in Table 7.



(a) Periocular (10% Same-session), Accuracy: 79.2%.

(b) Multi-faceted (10% Same-session), Accuracy: 80.4%.

(c) Periocular (10% Cross-session A), Accuracy: 52.2%.

(d) Multi-faceted (10% Cross-session A), Accuracy: 58.4%.

(e) Periocular (10% Cross-session B), Accuracy: 72.4%.

(f) Multi-faceted (10% Cross-session B), Accuracy: 90.5%.

Fig. 8. Confusion matrices of the recognition results for Subject 7 when fine-tuning by data with empirically estimated labels. These results correspond to the setting and performance reported in Table 7.

C Appendix: Interpreting Pupil Diameter Changes in Response to Emotions

Below, we provide an analysis of the recorded pupil diameter data, given its well-established association with emotional arousal in cognitive and affective computing research [11, 54, 104]. Specifically, we analyze changes in pupil diameter across different emotional categories in our dataset. We first pre-process the data to remove blinks, noise, and outliers. Next, we mitigate the influence of luminance on pupil diameter change, to isolate the pupil's response to emotional stimuli. Finally, we conduct both qualitative and quantitative analyses to examine emotion-specific patterns in pupil diameter changes.

The results reveal a clear trend: pupil dilation tends to increase for high-arousal emotions (e.g., fear) compared to neutral emotional states, which aligns with findings reported in prior literature [11, 54]. This analysis illustrates the potential of our dataset for studying emotion-induced physiological responses and serves as a representative example for future in-depth analysis of other biometric features [104] such as gaze direction, blinking rate, and periocular dynamics.

C.1 Eliminating Influence from Ambient Light

First, we apply the data pre-processing steps introduced in Section 5.2 to remove noise and outliers in the raw pupil diameter signals. However, pupil diameter is not only affected by changes in emotional state but is also strongly influenced by ambient illumination levels [16, 135]. In brighter environments, pupils constrict; in dimmer conditions, they dilate. Since our data is recorded from subjects wearing a VR headset, all ambient light came from the videos displayed within the headset. Thus, the illumination level can be retrieved for each frame of the videos by converting the frames of the videos into HSV color format and isolating the 'V' component which represents the luminance level. Given that the luminance source is known, its influence on pupil diameter can be modeled as [122]:

$$PD = PD_{\mathrm{L}} + PD_{\mathrm{E}}, \tag{1}$$

where, PD is the recorded pupil diameter, composed of a luminance-dependent component (PD_{L}) and an emotion-dependent component (PD_{E}). Following current studies [113, 122], we leverage a linear regression model to estimate the luminance-dependent component of the pupil diameter by:

$$PD_{\mathrm{L}} = a \times LV + b, \tag{2}$$

where LV is the luminance level retrieved from the video frame, a is the scaling factor and b represents the offset. Specifically, the coefficients a and b are obtained by fitting a linear regression between the average pupil diameter of both eyes recorded for the given participant, and the luminance value calculated for each movie frame [113, 122]. Finally, PD_{L} is subtracted from the recorded pupil diameter PD to obtain the emotion-dependent component PD_{E}. An example of the processed PD_{E} is shown in Figure 9, in which the original signal is recorded when the participant is watching the video "Still Alice". The estimated luminance-dependent component of pupil diameter, PD_{L} (in orange), has been subtracted from the recorded data PD (in blue) to obtain the component primarily influenced by emotional state, PD_{E} (in green).

C.2 Qualitative Analysis of Pupil Dilation and Constriction in Video Clips

Below, we present a case study that leverages the estimated emotion-dependent pupil diameter to qualitatively examine how pupil size changes in response to visual stimuli associated with different emotional states.

We take the session when one subject is watching the video clip "Still Alice" as an example. The video is selected to evoke the emotion of sadness (as shown in Table 9). In this scene, a woman and her husband are spending time at their beach house after learning she has early-onset Alzheimer's disease. Before going for a run, she attempts to find the bathroom but becomes disoriented, opening several doors without success. Eventually, her husband finds her standing in the hallway, having wet herself. She begins to cry, overwhelmed by the realization that she can no longer remember even simple things.

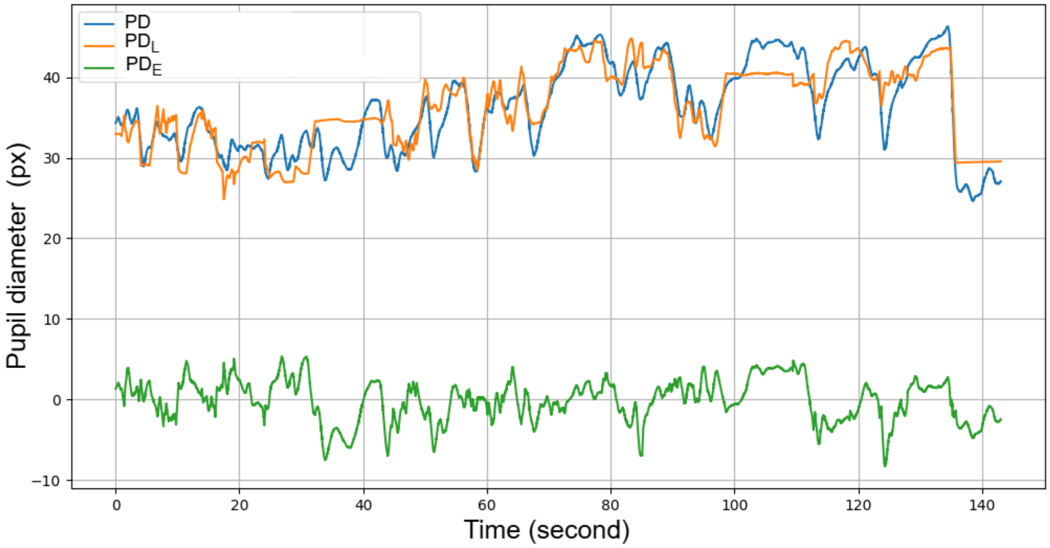


Fig. 9. An example of the emotion-dependent pupil diamter, PD_{E}, calculated by subtracting the luminance-dependent component PD_{L} from the recorded raw pupil diameter signal PD.

Happiness triggers pupil dilation. As shown in Figure 10, we annotate a few signal peaks observed in PD_{E}, which indicate moments of pupil dilation when the subject is watching the video. The four peaks are labeled numerically from 1 to 4. For each peak, we include a snapshot from the corresponding video segment along with a brief description of the associated scene. Peaks 1, 2, and 3 correspond to moments when: (1) the woman looks at old photos; (2) smiles at her newly arrived husband; and (3) when her husband appears with an amicable expression. These scenes evoked feelings of joy and happiness in the subject (as reported by the subject during data labeling), which also triggered pupil dilation. This aligns with previous studies [74, 80], which show that pupils dilate in response to pleasant stimuli such as images of loved ones, attractive faces, and emotionally engaging scenes. However, Peak 4 occurs when the woman becomes visibly distressed after forgetting where the bathroom is, and the subject's pupil dilates. This response was initially unexpected, since sadness has been linked to reduced pupil dilation in some studies [9, 11]. Our understanding is that the subject experienced empathy rather than sadness only. While empathy is not classified as one of the basic emotions, it often involves both cognitive and emotional engagement, increases arousal, and has been shown to cause pupil dilation [18, 130].

Sadness triggers pupil constriction. Similarly, in Figure 11, we annotate six signal troughs in PD_{E}, which indicate moments of pupil constriction. These troughs correspond to emotionally significant, sadness-inducing scenes. Specifically, Trough 1 occurs when the woman forgets she had just agreed to go running and continues looking at photos, prompting her husband to repeat the question that offers the first indication of her cognitive decline. Trough 2 occurs as she enters the house and immediately appears disoriented, revealing a second sign of her condition. Trough 3 aligns with a moment of clear confusion, intensified by the sad background music. Trough 4 occurs when the woman, visibly distressed, frantically opens several doors in search of the bathroom. In Trough 5, it is revealed that she has peed her pants. Finally, Trough 6 marks the moment she breaks down in tears as her husband approaches to comfort her. All six troughs coincide with sad moments and are consistent with typical pupil constriction observed during low-arousal emotional states such as sadness [9, 11].

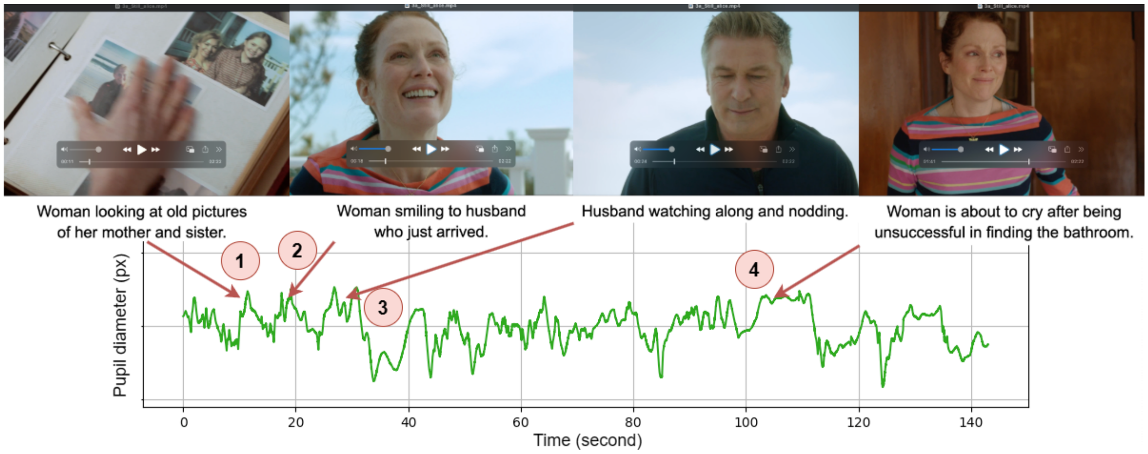


Fig. 10. Analyzing the signal peaks appear in PD_E and the corresponding scenes that elicit the subject's emotion.

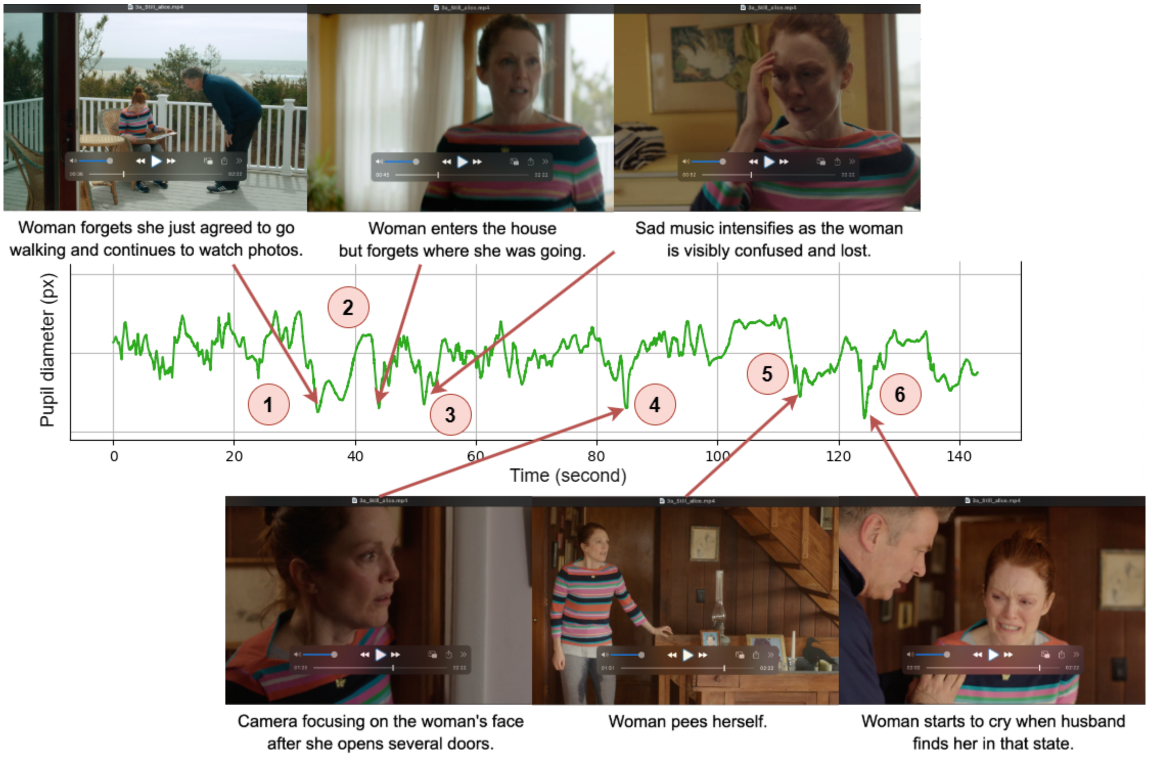


Fig. 11. Analyzing the signal troughs in PD_E and the corresponding scenes.

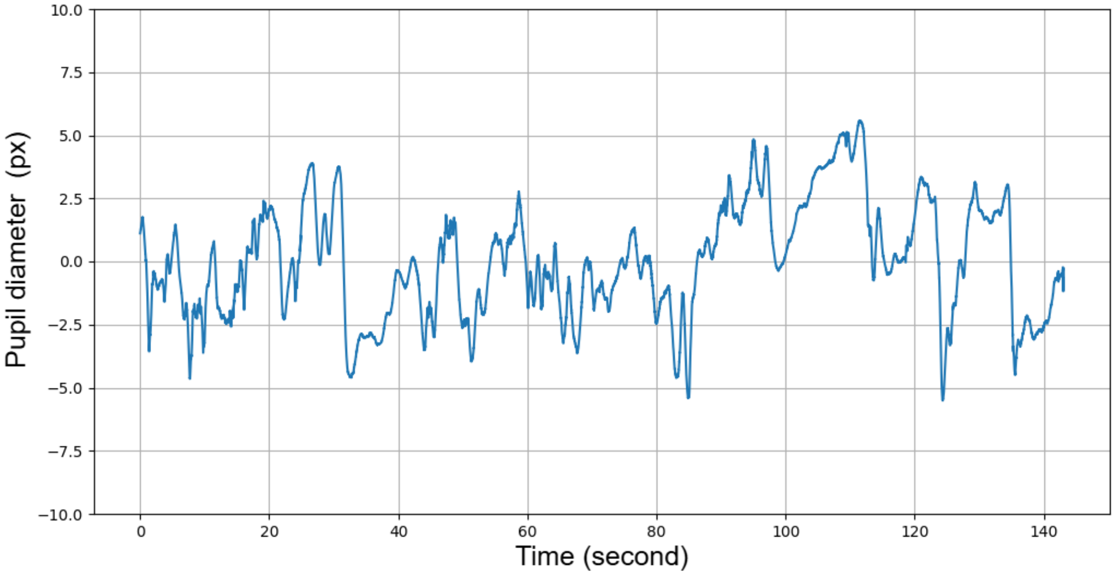


Fig. 12. The averaged PD_E over all subjects when watching the video stimulus “Still Alice". All subjects share similar overall patterns in pupil diameter changes, when comparing to that shown in Figure 9.

Overall investigation. Figure 12 shows the average PD_E across all subjects for the video “Still Alice”. Notably, the Peaks and Troughs observed in the participant-specific signal (Figure 9) are also appeared in the averaged signal. This observation suggests that subjects exhibit similar emotional response patterns throughout the video.

We apply the same processing steps to all recorded pupil diameter signals across the seven emotional states and computed the averaged PD_E for all participants corresponding to each emotion. The resulting distributions are shown in Figure 13. Note that, for clearer visual comparison, the box plots have been slightly shifted downward so that the median pupil diameter for the Neutral condition aligns with the x-axis. Moreover, as discussed earlier, that labeling an entire video clip with a single emotion can be inaccurate. Therefore, for each video, only the pupil data from the participants labeled segments were included in this analysis. In the box plots shown in Figure 13, the average PD_E values for Surprise, Happiness, Anger, Disgust, and Fear are higher than those for Neutral and Sadness. This aligns with findings from the recent study [122] that low-valence, low-arousal emotion (i.e., Sadness) lead to the greater pupil constriction, whereas high-valence, high-arousal emotions (i.e., Surprise) lead to the greater dilation.

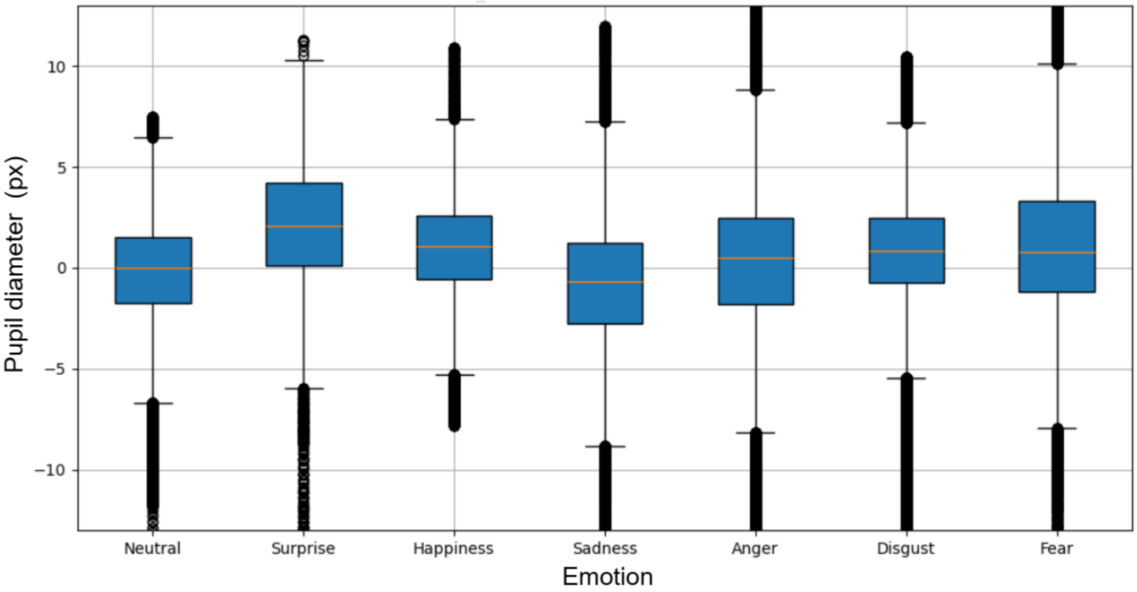


Fig. 13. The box plot shows the distribution of the PD_{E} of all participants combined, for each of the seven emotions. The low-valence and low-arousal emotion (i.e., Sadness) has greater pupil construction, whereas, high-valence and high-arousal emotion (i.e., Surprise) leads to the greater dilation.

D Appendix: Details of Hardware and Software Designs

This section gives the details of the hardware and software designs.

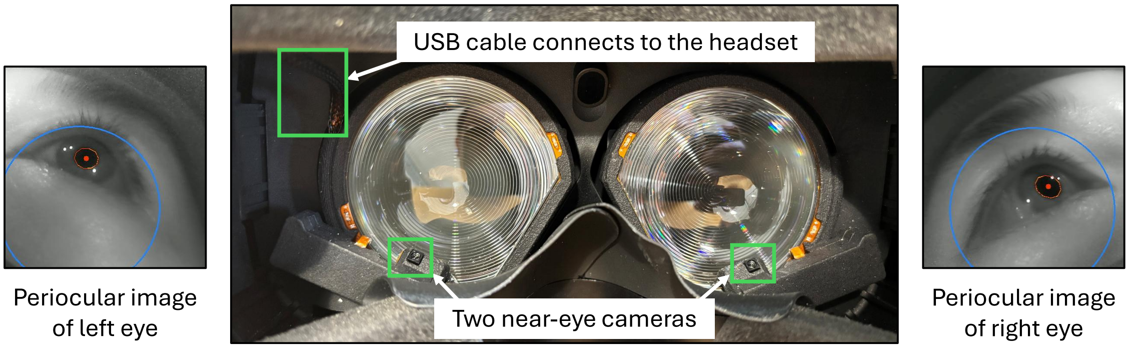

USB cable connects to the headset
Periocular image of left eye
Two near-eye cameras
Periocular image of right eye

Fig. 14. The hardware setup. The Pupil Labs eye-tracking add-on is connected to the VR headset through the USB port, which enables stable and interference-free video streaming of the periocular video.

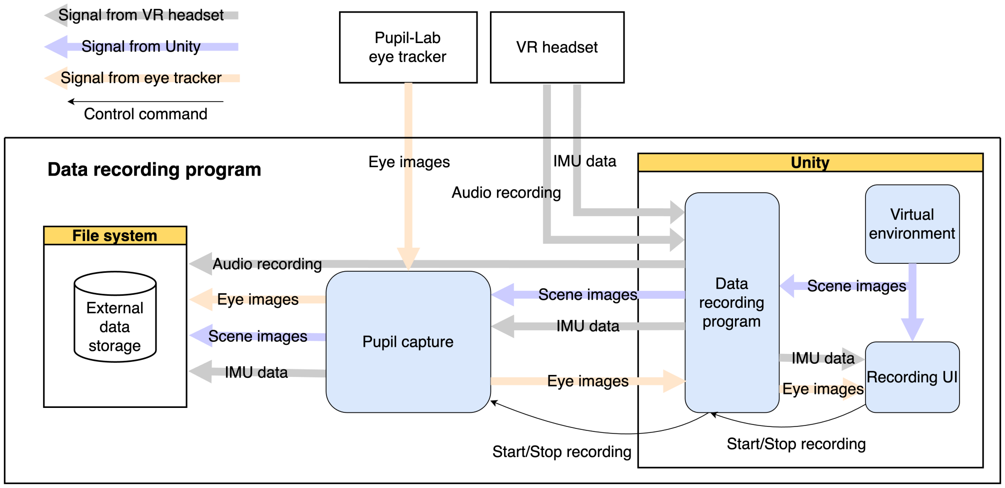


Fig. 15. The data recording program captures multiple signal streams from the VR headset and the Pupil Labs eye tracker. Specifically, four types of signals are recorded and stored in external storage: audio and IMU data from the VR headset's onboard sensors, near-eye images from the Pupil Labs eye tracker, and scene images generated in Unity, representing the participant's view in the virtual environment while watching the rendered stimuli.

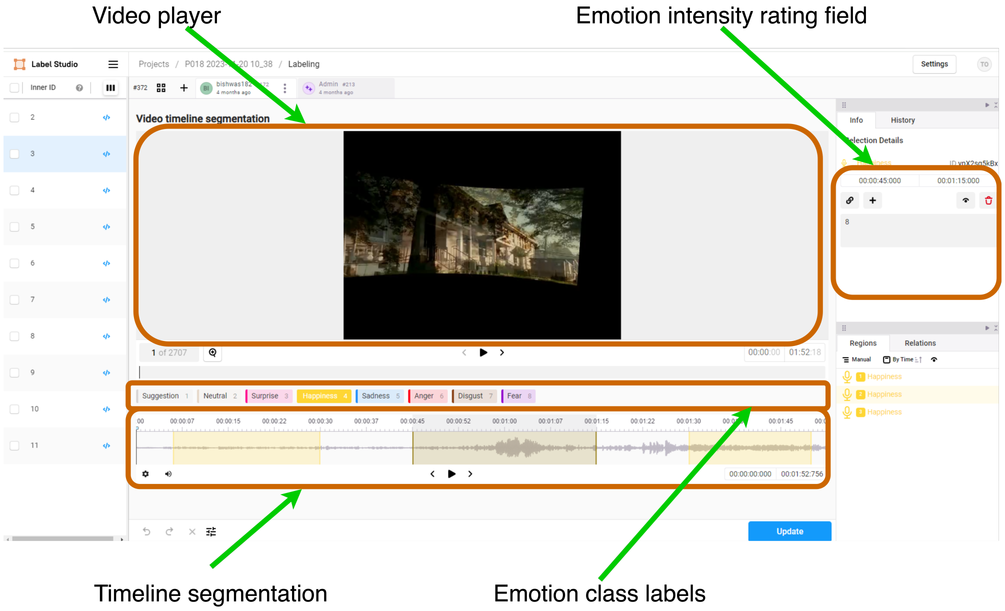


Fig. 16. An illustration of the Data Labeling UI and its main elements. After viewing each video stimulus, the participant, together with the researcher, can replay the video to create segments along the timeline and annotate each segment with an emotion label. Additionally, in the emotion intensity rating field, a numerical rating is assigned to each segment to represent the intensity of the participant's emotional response.