An IEEE Societies and Technical Councils Publication
IEEE Data Descriptions

Received 4 February 2025; revised 11 March 2025; accepted 16 March 2025; Date of publication 19 March, 2025; date of current version 3 April, 2025.

Digital Object Identifier 10.1109/IEEEDATA.2025.3553097

Descriptor: *Multimodal Dataset for Player Engagement Analysis in Video Games (MultiPENG)*

AMMAR RASHED [1], SHERVIN SHIRMOHAMMADI [1] (FELLOW, IEEE), AND MOHAMED HEFEEDA [2] (SENIOR MEMBER, IEEE)

[1]University of Ottawa, Ottawa, ON K1N 6N5, Canada
[2]Simon Fraser University, Burnaby, BC V5A 1S6, Canada

CORRESPONDING AUTHOR: Ammar Rashed (e-mail: arasi005@uottawa.ca).

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) through project Next Generation Cloud Gaming under Grant ALLRP556311-20.

ABSTRACT Player engagement is crucial for understanding and optimizing gaming experiences, yet the research community lacks comprehensive multimodal datasets with reliable engagement annotations. We present a dataset combining six synchronized data streams—EEG, eye tracking, heart rate, user inputs, webcam footage, and gameplay frames—collected from 39 participants playing popular games across varying difficulty levels. Our dataset's distinctive feature lies in its temporal precision, achieved through strategic integration of engagement surveys during natural game pauses, minimizing both recall bias and gameplay disruption. The dataset includes 900 annotated gameplay sessions with four psychological metrics (engagement, interest, stress, and excitement). Initial analyses revealed surprising findings: human judges achieved only 0.48 F1-score in engagement assessment from webcam footage, while a flow theory-based model reached 0.60 F1-score using difficulty and player experience. Our multimodal neural model combining EEG, eye tracking, and facial features demonstrated the dataset's potential with a 0.51 F1-score despite class imbalance. This comprehensive dataset enables various research directions in engagement measurement and modeling, supporting the development of more robust real-time engagement detection systems.

IEEE SOCIETY/COUNCIL Instrumentation and Measurement Society (IMS)

DATA TYPE/LOCATION Videos, Keystrokes, Physiological Signals

DATA DOI/PID 10.34740/kaggle/ds/6552328

INDEX TERMS Multimodal data of gameplay, player engagement measurement, psychological and physiological multisensory game data.

BACKGROUND

Player engagement measurement has emerged as a crucial component in gaming research, encompassing cognitive, emotional, and behavioral dimensions that necessitate comprehensive data collection approaches [1], [2]. While various engagement measurement techniques exist, from physiological sensors to behavioral analytics, the gaming research community has lacked publicly available multimodal datasets that combine these approaches with reliable ground truth annotations [3].

Our dataset addresses this gap by providing synchronized multimodal data streams from 39 participants playing two popular game titles: FIFA'23 and Street Fighter V (SFV). The dataset combines six key measurement modalities: heart rate data from a Fitbit smartwatch, 14-channel EEG recordings from an EPOC X headset, comprehensive eye-tracking metrics from a Gazepoint GP3 tracker, user input patterns from an Xbox controller, webcam footage, and gameplay frames. This multimodal approach enables researchers to study the relative effectiveness of different measurement

techniques and develop more robust engagement estimation methods.

Existing datasets have primarily focused on single modalities or limited combinations. The FaceEngage dataset [4] demonstrated the value of facial expressions for engagement detection but relied solely on game status for ground truth labels. The EngageMon dataset [5] showed the potential of sensor fusion in mobile gaming contexts but was limited in scope. More recently, player engagement estimation has evolved with datasets such as the Division 2 corpus [6], which demonstrated 72% accuracy in predicting long-term engagement using game footage and controller inputs, and GameVibe [7], which provided annotated gameplay sessions across 30 diverse games with third-person effect traces. These advancements align with emerging challenges and opportunities identified in comprehensive reviews of the field [8]. While datasets such as RECOLA and SEWA [9], [10] established protocols for synchronized multimodal collection, they were not gaming-specific.

Our dataset distinguishes itself through several key features. First, it employs the experience sampling method (ESM) [11] to collect ground truth engagement levels during natural game pauses, minimizing both gameplay disruption and recall bias. Second, it captures engagement across varied gameplay scenarios and difficulty levels, enabling analysis through flow theory [12] where player skill and game challenge interact. Third, it complements self-reported engagement with third-party annotations from trained judges, providing both subjective and objective perspectives on engagement indicators.

The dataset comprises 900 micro-game sessions from 39 participants (30 male, 9 female, mean age 24.3 years), with sessions distributed across both FIFA'23 and SFV. As shown in Table I, the dataset captures a wide range of engagement levels and related psychological states (interest, stress, and excitement), with session durations varying significantly between games (FIFA: mean = 91.5 s, SD = 50.3 s; SFV: mean = 36.7 s, SD = 9.8 s). The comprehensive nature of this dataset, combining multiple modalities with fine-grained temporal alignment and varied gameplay scenarios, provides researchers with rich opportunities for investigating player engagement across different game genres, difficulty levels, and measurement approaches.

COLLECTION METHODS AND DESIGN

Our data collection system integrates multiple specialized hardware and software components to capture diverse signals indicative of player engagement. The complete experimental setup, illustrated in Fig. 1, consists of six key components synchronized through a central gaming PC that serves as the primary data collection and synchronization hub. The setup is designed to maintain participant comfort while ensuring reliable data collection across all modalities.

A written consent was signed and obtained from all participants and the methodology was approved by the University of Ottawa's Office of Research Ethics and Integrity, under file number H-07-23-9439.

TABLE I. Session Counts and Durations (Minutes) per Game and Dimension

Dimension	Level	FIFA		SFV		Total	
		#	Dur.	#	Dur.	#	Dur.
Engagement	0–1	14	16	88	47	102	63
	2	26	40	137	79	163	119
	3–4	97	170	538	342	635	512
Interest	0–1	6	10	90	48	96	58
	2	56	88	227	134	283	222
	3–4	75	127	446	287	521	414
Stress	0–1	53	71	282	164	335	235
	2	40	74	264	165	304	239
	3–4	44	81	217	140	261	221
Excitement	0–1	36	50	194	108	230	158
	2	56	102	294	182	350	283
	3–4	45	74	275	179	320	253
Total		137	226	763	468	**900**	**694**

Note: The overall size of dataset are highlighted in bold.

FIG. 1. Experimental setup showing the integrated hardware components: 1) webcam; 2) survey tablet; 3) eye tracker; 4) smartwatch; 5) gamepad; and 6) EEG headset. ©ACM reused with permission from [8]/cropped and horizontally flipped from original and annotated.

Hardware Configuration

The primary data acquisition hardware includes a 1080p webcam mounted on the monitor for capturing facial expressions and head pose, an EPOC X EEG headset operating at 128 Hz with 14 channels for brain activity measurement, and a Gazepoint GP3 Eye Tracker positioned below the monitor sampling at 60 Hz. Participants wear a Fitbit Versa 3 smartwatch on their left wrist for heart rate monitoring, while gameplay input is captured through an Xbox USB gamepad. A separate touchscreen tablet is positioned for engagement survey responses, and an additional screen is used to monitor the experiment through a unified interface.

The experiment environment is carefully controlled to simulate a natural gaming setting while maintaining data quality. The testbed operates in a sound-isolated room with

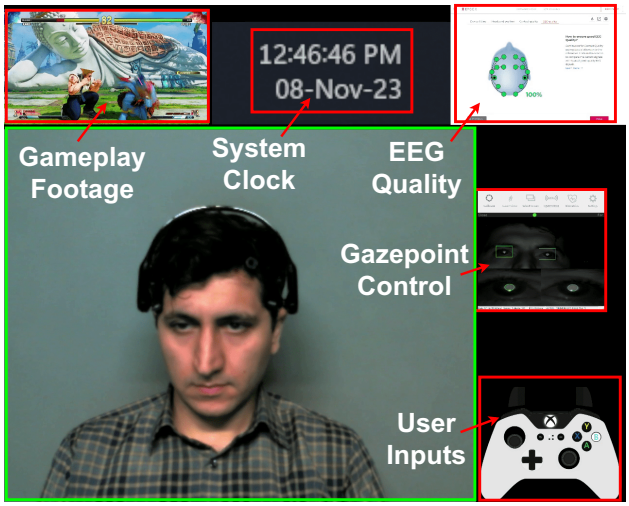

FIG. 2. OBS recording scene showing the synchronized capture of webcam feed, gameplay footage, and system metrics. This unified view provides a comprehensive record of each session and enables real-time quality monitoring.

consistent lighting to minimize variations in webcam and eye-tracking data. The viewing distance and monitor angle are standardized (65 cm from eye tracker, monitor tilted at 15°) to maintain eye-tracking accuracy across sessions. An adjustable chair ensures participant comfort throughout the session.

Software Infrastructure

The software infrastructure consists of several specialized components operating in concert. OBS Studio captures both the webcam feed and gameplay footage, with audio recorded from the webcam's built-in microphone and the game's audio output. As shown in Fig. 2, the OBS interface serves as a unified auditing system, displaying the webcam feed, gameplay footage, system clock, EEG signal quality (SQ), and eye-tracking quality metrics in a single view. This synchronized display enables real-time monitoring of data quality across all modalities and provides a comprehensive record of each session.

The collected webcam footage was processed using OpenFace [13], a comprehensive facial behavior analysis toolkit. This processing extracted detailed facial features including head pose dynamics (translation, rotation, velocity, and acceleration vectors), facial landmarks, gaze direction estimates, and 17 facial action unit (AU) intensities. These facial analysis capabilities complement the dedicated eye-tracking hardware while providing additional metrics such as head movement patterns that have been shown to correlate with engagement in prior work [4].

The Gazepoint analysis software captures comprehensive eye metrics including gaze position relative to the screen, pupil dilation, fixations, saccades, and blink rate. The Gazepoint control software manages device calibration and maintains tracking accuracy throughout the session. The Emotiv Pro software handles real-time EEG signal acquisition and processing, providing both raw EEG signals and derived performance metrics including attention, engagement, and stress levels. SQ is continuously monitored through contact quality (CQ), machine learning SQ, and signal magnitude quality (SMQ) metrics.

Heart rate data are logged and uploaded in real-time to the Google account connected to the smartwatch at 5-s intervals, while a custom Python script handles gamepad input logging, recording timestamped controller interactions including button presses (A,B,X,Y, bumpers) and stick positions (left/right X/Y coordinates from $-32\,767$ to $32\,767$).

Data Collection Protocol

The collection process begins with participant registration, capturing demographic information and gaming experience, particularly familiarity with soccer and 2-D fighting games. The calibration sequence includes the eye tracker's standard nine-point calibration procedure using Gazepoint control, followed by a 15-s eyes-open and eyes-closed EEG baseline recording. The Fitbit is secured snugly on the left wrist, and proper electrode contact is verified for the EEG headset using saline solution to ensure optimal SQ.

Two popular games were strategically selected based on their representativeness, accessibility, difficulty variability, and natural session boundaries. FIFA'23 and SFV represent two distinct and popular gaming genres (sports and fighting) with different gameplay paces and skill requirements. These games offer precise difficulty control (FIFA: 6 levels, SFV: 8 levels) allowing systematic investigation of engagement across challenge intensities. Importantly, while FIFA typically requires prior experience for meaningful play, SFV's simple core mechanics enabled participation from players with minimal gaming background while still offering depth through advanced techniques. This complementary selection ensured our dataset captures a broader spectrum of player experiences and skill levels than would be possible with a single game type.

Both games feature natural pauses that facilitate nondisruptive survey administration—a critical design consideration for maintaining ecological validity. For FIFA'23, only participants with prior soccer gaming experience participated, playing 3–5 matches with surveys conducted after goals, at half-time, and post-match. To prevent survey fatigue while maintaining data quality, a minimum 20-s gameplay duration is enforced between consecutive surveys.

For SFV, participants undergo a 5–10-min training phase until they report comfort with basic controls and mechanics. Each round has a maximum duration of 99 s, though rounds typically conclude earlier through knockouts. Participants are targeted to play three rounds at each difficulty level (low: 1–3, medium: 4–5, and high: 6–8), with the actual number varying based on remaining session time and training duration. Surveys are administered between rounds, coinciding with the game's natural break points.

TABLE II. Survey Questions and Response Options

Dimension	Question and Response Scale	
Engagement	How engaged did you feel?	
	0. Very Bored	1. Somewhat Bored
	2. Neutral	3. Somewhat Engaged
	4. Very Engaged	
Interest	How much did you enjoy?	
	0. Strongly Disliked	1. Disliked
	2. Neutral	3. Liked
	4. Strongly Liked	
Stress	How stressed did you feel?	
	0. Very Relaxed	1. Relaxed
	2. Somewhat Stressed	3. Stressed
	4. Very Stressed	
Excitement	How excited did you feel?	
	0. Not Excited	1. Slightly Excited
	2. Moderately Excited	3. Extra Excited
	4. Extremely Excited	

TABLE III. EEG Data Structure

Category	Rate	Column Format	Values
Raw EEG	128 Hz	EEG.channel	μV readings
Contact Quality	128 Hz	CQ.Overall	0–100
		CQ.channel	0–4
Signal Quality	2 Hz	EQ.Overall	0–100
		EQ.channel	0–4
Performance Metrics	0.1 Hz	PM.metric	Type: - IsActive (0/1) - Scaled (0–1) - Raw (unbounded) - Min, Max (bounds)
Band Powers	8 Hz	POW.channel	Band type: - Theta (4–8 Hz) - Alpha (8–12 Hz) - BetaL (12–16 Hz) - BetaH (16–25 Hz) - Gamma (25–45 Hz)

Survey Design

The survey application captures self-reported metrics across four key dimensions using five-point Likert scales, as detailed in Table II. The selection of these specific metrics serves two theoretical frameworks. First, while engagement serves as the primary metric, interest, and excitement map to the fundamental dimensions of emotion measurement (valence and arousal), while interest connects to conation—the desire to continue playing—which is often used as an engagement proxy [14]. Stress levels relate to the flow theory of engagement [12], particularly regarding game challenge intensity. Second, these metrics mirror those reported by the EMOTIV Pro EEG software but in a gaming-specific context, enabling analysis of correlations between general EEG-based metrics and gaming-specific self-reported states. The engagement dimension ranges from very bored (0) to very engaged (4). Interest is measured from strongly disliked (0) to strongly liked (4). Stress levels span from very relaxed (0) to very stressed (4), and excitement ranges from not excited (0) to extremely excited (4).

Synchronization Implementation

The synchronization system aligns all data streams through careful clock calibration. The high-frequency data streams (EEG at 128 Hz, eye tracking at 60 Hz, and gameplay footage at 30 fps) are logged on the gaming PC with a single clock. While the smartwatch clock exhibits a 1–2-s gap with the PC clock, this discrepancy is tolerable given its 5-s sampling interval for heart rate data. The PC and survey tablet clocks are manually calibrated to the smartwatch's clock with subsecond discrepancy, ensuring temporal alignment across all data streams while accommodating the lower sampling rate of the heart rate measurements.

VALIDATION AND QUALITY

To validate the quality and utility of our dataset, we present evidence supporting both our measurement framework and demonstrate the dataset's effectiveness through multiple use cases.

Quality Monitoring

We collected quality metrics for EEG, heart rate, and eye-tracking samples provided by the corresponding data collection software. The EPOC X EEG headset provides continuous SQ metrics including CQ, machine learning SQ, and SMQ. These quality metrics are aggregated into a 0–100 overall quality score indicated in column EQ.OVERALL (see Table III). The Fitbit heart rate measurements include confidence levels (0–3 scale). We also use the FPOGV flag in gazepoint data as a binary quality score indicating whether there is a valid point of gaze (POG) detected. These metrics are included in the dataset, allowing researchers to establish appropriate quality thresholds for their specific analyses.

To understand the effect of different quality thresholds on the dataset size, we calculate the average quality scores per sample (i.e., an annotated game session) for each of the three modalities. Fig. 3 shows the complementary cumulative distribution function (CCDF) of samples given a quality threshold. Interestingly, only 50% of samples have an average EEG quality (EQ.OVERALL) of at least 75%. This is mostly due to sudden player movements during gameplay, which we observed to cause a short-term decline in EEG SQ. This can be an interesting research direction to explore the relationship between sudden drops in EEG quality as a proxy for sudden movement and highly engaging gameplay moments. Similarly, the heart rate signals show relatively low confidence overall. These results emphasize the importance of postprocessing physiological signals to maximize their utility in modeling player engagement. Most eye-tracking

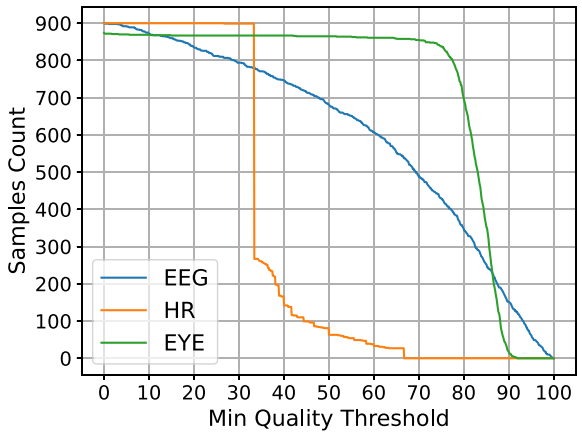

FIG. 3. **Sample survival rate versus score threshold.**

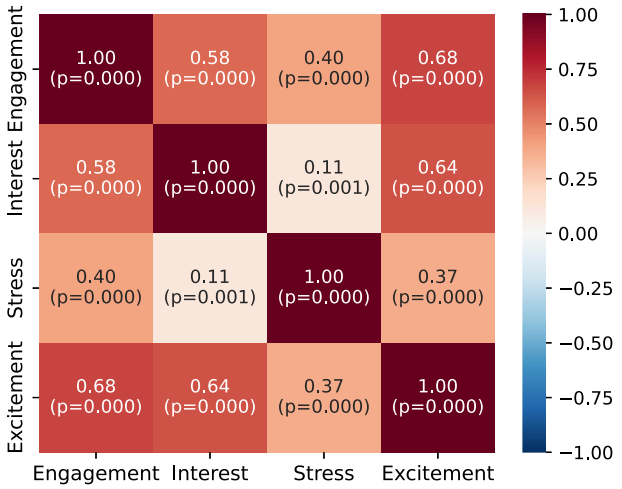

FIG. 4. **Spearman correlation matrix between engagement metrics.**

samples have 75%–80% valid POG. Notably, invalid FPOGV flag are often associated with blinking events, which do not necessarily indicate noisy data.

Technical Limitations and Dataset Scope

While our dataset offers rich multimodal data across different games and participants, we acknowledge certain constraints in its scope and technical implementation. The game selection, though strategically chosen to represent different genres and skill levels, is limited to two titles and may not fully represent all gaming experiences across the vast landscape of game genres. However, this limitation is partially mitigated by our diverse participant demographics, which include individuals ranging from those with no prior gaming experience to casual players and experts in the selected games.

Prior to the main data collection, we conducted an initial trial with four participants to refine the experimental protocol and identify potential issues. During the full data collection, technical and procedural limitations were encountered.

1) One participant (ID: 559) reported being previously diagnosed with ADHD and was using stimulant medication during the experiment. The divergence between this participant's EEG signals and the baseline signal resulted in a diminished EEG quality score, consistent with prior research on stimulant effects on EEG signals [15].
2) Six participants (IDs: 872, 850, 568, 533, 297, and 183) were recorded under different lighting conditions than the standard protocol.
3) Controller input data were not captured for ten participants (IDs: 120, 166, 462, 539, 623, 703, 754, 507, 514, and 744).
4) Eye-tracking data were incomplete for one participant (ID: 407).

These missing data points represent a limitation that researchers should consider when analyzing the affected sessions. However, our dataset's modular structure enables researchers to selectively include participants based on available modalities for specific research questions. For instance, in our multimodal neural architecture implementation, we utilized only participants with complete data for the specific modalities being investigated. The multimodal nature of our dataset provides inherent redundancy across different data streams, potentially enabling more robust analyses even when certain modalities are unavailable for some participants. We recommend that researchers clearly document which participant subsets they use for each analysis to ensure reproducibility. Future extensions of this work could include additional game genres and address these technical challenges to build upon the foundation established by this dataset.

Validation of Engagement Dimension Selection

The theoretical framework underlying our 4-D survey design was validated through two complementary analyses. First, correlation analysis between metrics revealed meaningful relationships supporting our measurement approach, as shown in Fig. 4. Engagement showed strong positive correlations with excitement ($\rho = 0.68$, $p < 0.001$) and interest ($\rho = 0.58$, $p < 0.001$), validating our connection to fundamental dimensions of emotion measurement (valence and arousal). The moderate correlation between stress and engagement ($\rho = 0.40$, $p < 0.001$) aligns with flow theory's emphasis on challenge intensity, while the weak correlation between interest and stress ($\rho = 0.11$, $p = 0.001$) confirms these capture distinct aspects of gameplay experience.

Second, comparison with EMOTIV Pro's EEG-based metrics during gaming sessions revealed important insights about engagement measurement in gaming contexts. The weak correlation between EEG-measured engagement and self-reported engagement ($\rho = 0.076$, $p < 0.05$), along with similarly weak correlations for other metrics (EEG-measured stress: $\rho = 0.089$, $p < 0.01$; interest: $\rho = 0.003$, $p = 0.92$; excitement: $\rho = -0.113$, $p < 0.001$), validates our choice of gaming-specific engagement dimensions over general EEG-based metrics. These results demonstrate that while

commercial EEG systems can measure general cognitive states, gaming engagement requires domain-specific measurement approaches.

RECORDS AND STORAGE

The *Multimodal Player Engagement* dataset is publicly available on Kaggle[1] and is organized as follows.

```
Samples/
├── pid /
│   ├── EEG/
│   ├── EYE/
│   ├── HR/
│   ├── OBS/
│   ├── OpenFace/
│   └── XBOX/
├── Questionnaires/
│   ├── participants.csv
│   └── submissions.csv
├── splits/
│   ├── fold_[0-6]/
│   │   └── [0-6].csv
Human Panel Samples/
├── Samples/
│   └── sid.mp4
├── Training/
├── annotations.csv
└── annotators.csv
```

The *Samples* folder contains the primary multimodal data collection organized by participant (`pid`: participant_id) and modality. The *Questoinnaires* folder contains the primary ground truth data collected during the experiment through participant surveys. The *splits* folder provides a standardized benchmark framework implementing nested stratified group cross validation. The *Human Panel Samples* folder provides a curated subset of gameplay sessions for human evaluation of player engagement through visual cues. We explain each folder in the following.

Samples: The Multimodal Data

Each participant's data are stored in a separate subfolder identified by their unique participant ID (`pid`), with further subfolders for each data modality (EEG, eye tracking, heart rate, OBS scene, and XBOX controller inputs). The hierarchical structure ensures clear organization of the extensive multimodal data while maintaining the relationship between different data streams for each gaming session. The naming convention of samples is `pid_sid_eng_int_str_exc`, where `sid`: submission_id, `eng`: engagement, `int`: interest, `str`: stress, and `exc`: excitement. The file extension of samples is `.mp4` for OBS videos, and `.csv` for the rest.

[1] kaggle.com/datasets/ammarrashed23/multimodal-player-engagement

EEG Data Files

The EEG files contain measurements from an EPOC X headset with 14 channels (AF3, AF4, F3, F4, F7, F8, FC5, FC6, O1, O2, P7, P8, T7, and T8). Each file begins with metadata columns: a timestamp in datetime format (e.g., "2023-10-31 14:24:22.025-04:00"), a sample Counter (0-127, resets every second), and an Interpolated flag indicating whether the sample was received from the headset (0) or interpolated. The data are organized into several measurement categories, each sampled at different frequencies, as shown in Table III. The primary data consist of raw EEG voltages sampled at 128 Hz. For each channel, the SQ is monitored through two metrics: CQ indicates the physical connection quality between electrodes and scalp, while EEG quality (EQ) provides a more comprehensive SQ assessment updated every 500 ms. The headset computes higher level performance metrics at 0.1 Hz, including measures of engagement, excitement, stress, relaxation, interest, and focus. Each metric includes both raw algorithm outputs and normalized values. Additionally, the power in five frequency bands is computed for each channel at 8 Hz, providing insights into different aspects of brain activity during gameplay.

Eye-Tracking Data Files

The eye-tracking data are collected at 60 Hz using a Gazepoint GP3 eye tracker. Each record contains a timestamp and a video frame counter (VID_FRAME), along with three main categories of measurements.

1) The gaze data include both filtered (FPOG) and unfiltered (BPOG) point-of-gaze coordinates. Filtered coordinates (FPOGX, FPOGY) represent fixation points with associated start time (FPOGS), duration (FPOGD), and a unique identifier (FPOGID). Unfiltered coordinates (BPOGX, BPOGY) provide raw gaze positions. Each measure includes a validity flag (FPOGV, BPOGV).
2) Individual eye measurements track pupil position (LPCX/RPCX, LPCY/RPCY), diameter in both pixels (LPD/RPD) and millimeters (LPMM/RPMM), and a scale factor normalized to the calibration depth (LPS/RPS). Each measurement includes its validity flag (LPV/RPV, LPMMV/RPMMV).
3) The system also tracks blink events with unique identifiers (BKID), durations (BKDUR), and frequency (BKPMIN, blinks per minute), as well as saccade characteristics including magnitude (SACCADE_MAG) and direction (SACCADE_DIR). A pixel-to-millimeter conversion scale (PIXS) is provided with its validity flag (PIXV).

Heart Rate Data Files

The heart rate data are collected at 0.2 Hz (every 5 s) using a Fitbit Versa 3 smartwatch. Each record contains a timestamp, heart rate in beats per minute (BPM), and a confidence measure (0–3) indicating the reliability of

TABLE IV. Structure of Questionnaire Files

File	Column	Description
participants.csv	participant_id	Unique identifier
	age	Participant age
	sex	M/F
	fifa_exp	FIFA experience (0–4)
	sf_exp	Street Fighter experience (0–4)
submissions.csv	submission_id	Unique session identifier
	participant_id	Player identifier
	game	FIFA23 or Street Fighter V
	difficulty	Game-specific level
	session_no	Sequential session number
	start_ts	Session start
	end_ts	Session end timestamp
	engagement	Overall engagement (0–4)
	interest	Interest/enjoyment (0–4)
	stress	Stress level (0–4)
	excitement	Excitement level (0–4)

the reading. The confidence value is determined by the smartwatch's internal algorithms based on factors such as sensor CQ and motion artifacts.

Controller Input Data Files

The Xbox controller inputs are recorded asynchronously (event-driven) and consist of two types of events: analog inputs (Absolute) and button presses (Key). Each record contains a timestamp and the event details, including the specific control (Event) and its state.

Analog inputs (EventType: "Absolute") include stick positions (left_stick_x, left_stick_y, right_stick_x, right_stick_y), trigger depths (left_trigger, right_trigger), and d-pad directions (dpad_x, dpad_y). For these events, the state ranges from −32 767 to 32 767, representing the full range of motion. Button events (EventType: "Key") capture binary states (0 or 1) for all controller buttons: face buttons (a_button, b_button, x_button, y_button), bumpers (left_bumper, right_bumper), stick clicks (left_stick_button, right_stick_button), and menu buttons (start_button, back_button).

Questionnaires: The Survey Data

This folder contains the primary ground truth data collected during the experiment through participant surveys. The data are organized in two CSV files shown in Table IV: a participant registry capturing demographic information and gaming experience, and a comprehensive session log containing engagement metrics and contextual information. All timestamps follow the format (YYYY-MM-DD HH:MM:SS-ZZZZ). For FIFA23, difficulty levels progress from easiest to most difficult as: *Beginner, Amateur, Semi-Pro, Professional, World Class, Legendary*. SFV difficulties are indicated numerically from *(1)–(8)*, where 1 is easiest and 8 is most difficult. The session number is reset for each unique participant-game-difficulty combination to track progression within specific difficulty levels.

TABLE V. Structure of Human Panel Folder

Folder/File	Contents/Column	Description
Samples/	20 files	Webcam footage cropped
	session_id.mp4	from OBS recordings
Training/	high1, high2	Engagement labels 3–4
	neutral1, neutral2	Engagement label 2
	low1	Engagement labels 0–1
annotations.csv	participant_id	Unique player identifier
	submission_id	Unique session identifier
	engagement	Ground truth rating (0–4)
	annotator_[0-13]	Annotator ratings (0–4)
	annotator_[0-13]_conf	Annotator confidence (0–4)
annotators.csv	annotator_id	Unique identifier (0–13)
	experience	Gaming experience (0–4)
	clues	Engagement indicators used

"Splits": The Cross-Validation Folds

The structure consists of seven outer folds, each is further divided into seven inner folds. Each test set contains 4–6 participants, while validation sets comprise 3–5 participants. The splitting strategy ensures representation of minority classes across all folds to address data imbalance concerns. This nested structure supports various evaluation approaches: the outer folds provide unbiased performance estimates, while inner folds enable systematic hyperparameter tuning or ensemble model development. The consistent participant-level splitting across all folds ensures reproducible benchmarking for future research using this dataset.

"Human Panel Samples": The Human Evaluation Subset

As detailed in Table V, the folder contains standardized webcam recordings and corresponding annotation data. The Samples folder contains 20 gameplay sessions, with video files named using the format `< session_id > .mp4`. Each video maintains consistent quality specifications (480 x 480 pixels, 30 FPS) achieved by cropping the player webcam feed from the original OBS recordings. The Training folder contains five reference samples representing distinct engagement levels (two high, two neutral, and one low), sourced from different participants to establish diverse baseline examples. These training samples were used to calibrate annotators and establish common rating criteria before their evaluation of the main sample set. The samples were selected to represent various engagement levels and player demographics, enabling comprehensive assessment of human annotators' rating consistency and accuracy. To facilitate evaluation against human annotators, the samples were sourced exclusively from the test set of the final outer fold (fold 6).

TABLE VI. Performance Comparison of Baselines

Model	Class	Precision	Recall	F1-score
Human*	High	0.38 ± 0.14	0.45 ± 0.22	0.40 ± 0.16
	Low	0.59 ± 0.13	0.53 ± 0.18	0.55 ± 0.13
Flow-based	High	0.77 ± 0.08	0.75 ± 0.08	**0.76** ± 0.05
	Low	0.43 ± 0.16	0.45 ± 0.14	**0.43** ± 0.12
Multimodal	High	0.73 ± 0.08	0.69 ± 0.19	0.70 ± 0.11
	Low	0.28 ± 0.20	0.36 ± 0.24	0.31 ± 0.22

Note: *Human evaluation performed on 20 random samples only. The bold values indicate the highest performance.

INSIGHTS AND NOTES

To demonstrate our dataset's suitability for developing engagement detection systems, we implemented three baseline approaches for real-time engagement prediction, as shown in Table VI. We formulated the task as binary classification, differentiating between engagement levels above and below our dataset's mean of 2.87. All experiments used sevenfold stratified group cross validation to ensure participant-level separation between training and testing data.

Human Annotation Analysis

First, we conducted human annotation analysis to understand the capabilities and limitations of humans in visually assessing engagement in gaming contexts, not as an alternative ground truth but as a baseline for comparison against sensor-based approaches. We recruited 14 human judges who, after training on the said five reference samples, analyzed webcam footage from 20 gameplay sessions across five different participants. The training consisted of providing judges with sample videos representing different engagement levels and suggesting potential visual cues (e.g., eye blinking patterns, head and eye movements, and facial expressions) without mandating specific indicators to observe. When asked about their rating criteria, judges reported primarily using eye blinking patterns, head and eye movements, and facial expressions as cues.

Inter-rater agreement analysis revealed consistently low agreement across different granularities: raw five-point Likert scores (Krippendorff's $\alpha = 0.092$, Cohen's $\kappa = 0.001$), low/neutral/high classes ($\alpha = 0.080$, $\kappa = 0.036$), and binary low/high classes (raw agreement: 52.1%, $\alpha = 0.040$, $\kappa = 0.043$). When evaluated against survey answers, annotators achieved $50\% \pm 12\%$ accuracy. This low agreement, consistent across multiple agreement metrics and classification granularities, highlights a fundamental challenge in engagement measurement: unlike more explicitly manifested emotional states (such as happiness indicated by smiling), engagement appears to lack consistently interpretable visual cues. This finding underscores why our dataset relies on self-reported engagement as ground truth rather than third-party observations, and why multimodal approaches that incorporate physiological and behavioral signals beyond visual cues are necessary. The limited accuracy of human judges using only visual information serves as an important baseline when comparing against our computational approaches that leverage multiple data streams. This finding aligns with previous research suggesting engagement is a complex internal state that may not reliably manifest in observable facial expressions or behavioral patterns.

Flow Theory-Based Engagement Detection

Second, building on flow theory, which suggests optimal engagement occurs when skill matches challenge levels, we developed a random forest classifier using participant experience as a skill proxy and normalized game difficulty as a challenge measure. The model incorporated two key features: player skill (represented by self-reported experience levels on a five-point scale) and normalized game difficulty (six levels for FIFA'23, eight levels for SFV, normalized to 0–1 range). This relatively simple approach achieved 66% accuracy (±6%) across the full dataset, with particularly strong performance in detecting high-engagement states (precision: 0.77 ± 0.08, recall: 0.75 ± 0.08). While effective, this method's applicability is inherently limited to games with clear difficulty levels and quantifiable player skill measures.

Multimodal Neural Architecture

Third, to explore the dataset's potential for comprehensive engagement prediction, we implemented a neural architecture combining three key modalities: EEG, eye tracking, and facial features. The input streams were processed as follows.

For EEG data, we used band powers (theta, alpha, low/high beta, and gamma) from all 14 channels as provided by the EMOTIV software without additional filtering or artifact removal beyond the system's built-in processing.

For eye tracking, we extracted specific features from the Gazepoint software output including fixation position and duration (FPOGX, FPOGY, and FPOGD), saccade characteristics (SACCADE_MAG, SACCADE_DIR), pupil diameter (LPD, RPD), and blink patterns (binary blink state, blink duration, and blinks per minute).

Facial features were extracted from OpenFace [13] analysis following the approach in [4], focusing on head pose dynamics (3-D translation vectors and their derived velocity and acceleration) and facial AU intensities. We utilized the intensity values (ranging from 0 to 5) of all 17 AUs detected by OpenFace, which include raising chin (AU17), stretching lips (AU20), and blinking (AU45).

Heart rate data, while collected and included in the dataset, were not incorporated into our neural architecture. This decision was influenced by the known challenges in processing photoplethysmography (PPG) signals from consumer-grade wearables. The complexity of properly filtering and normalizing these signals, especially during physical movement associated with gameplay, presents an opportunity for future work with this dataset.

The architecture employs modality-specific encoders with temporal convolutions (kernel size = 5, stride = 2) and

global pooling. Each encoder processes 5-s windows of its respective input stream. The EEG encoder handles 14×5 matrices (channels $\times$ frequency bands), the eye-tracking encoder processes 8-D feature vectors (fixation metrics, pupil size, and blink rate), and the facial feature encoder takes 20-dimensional vectors (AU intensities and head pose parameters).

Late fusion combines these encoded features through concatenation, followed by a two-layer MLP (hidden units = 128*3 then 128) with dropout (0.3) between layers. We used the Adam optimizer with learning rate $1e{-}4$ and a class-weighted cross-entropy loss function to address the class imbalance (635 high-engagement samples versus 265 low-engagement samples). Following our strict participant-level separation protocol, each fold's training data were further split using stratified group sampling, holding out 3–4 participants as a validation set for early stopping.

The multimodal approach achieved $61\% \pm 9\%$ accuracy, with notably strong performance in high-engagement detection (precision: 0.73 ± 0.08) but challenges in low-engagement cases (precision: 0.28 ± 0.20). This performance pattern reflects both the dataset's class imbalance and the inherent difficulty of detecting disengagement states. While the overall accuracy is lower than the flow-based approach, the multimodal model demonstrates the feasibility of engagement prediction using purely sensor-based inputs, without requiring game-specific knowledge such as difficulty levels.

SOURCE CODE AND SCRIPTS

The scripts used in this work are publicly available in the GitHub *AmmarRashed/MultimodalEngagement* repository.[2]

ACKNOWLEDGMENTS

A.R. collected, curated, and analyzed the data and wrote the first draft of the manuscript. S.S. and M.H. guided the experiment design, directed the data analysis, and provided feedback and edits. All authors reviewed the manuscript.

The authors have declared no conflicts of interest.

REFERENCES

[1] H. L. O'Brien, I. Roll, A. Kampen, and N. Davoudi, "Rethinking (dis) engagement in human-computer interaction," *Comput. Hum. Behav.*, vol. 128, 2022, Art. no. 107109.

[2] S. Poeller, S. Seel, N. Baumann, and R. L. Mandryk, "Seek what you need: Affiliation and power motives drive need satisfaction, intrinsic motivation, and flow in league of legends," *Proc. ACM Human-Comput. Interact.*, vol. 5, no. CHI PLAY, pp. 1–23, 2021.

[3] D.-I. D. Han, F. Melissen, and M. Haggis-Burridge, "Immersive experience framework: A Delphi approach," *Behav. Inf. Technol.*, vol. 43, no. 4, pp. 623–639, 2024.

[4] X. Chen, L. Niu, A. Veeraraghavan, and A. Sabharwal, "FaceEngage: Robust estimation of gameplay engagement from user-contributed (YouTube) videos," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 651–665, Apr./Jun. 2022.

[5] S. Huynh, S. Kim, J. Ko, R. K. Balan, and Y. Lee, "EngageMon: Multimodal engagement sensing for mobile games," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 1, pp. 1–27, 2018.

[6] K. Pinitas et al., "Predicting player engagement in Tom Clancy's the division 2: A multimodal approach via pixels and gamepad actions," in *Proc. 25th Int. Conf. Multimodal Interact., (ICMI)*, New York, NY, USA: ACM, 2023, pp. 488–497.

[7] M. Barthet, M. Kaselimi, K. Pinitas, K. Makantasis, A. Liapis, and G. N. Yannakakis, "GameVibe: A multimodal affective game corpus," *Sci. Data*, vol. 11, no. 1, 2024, Art. no. 1306.

[8] A. Rashed, S. Shirmohammadi, I. Amer, and M. Hefeeda, "A review of player engagement estimation in video games: Challenges and opportunities," *ACM Trans. Multimedia Comput., Commun. Appl.*, 2025.

[9] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Shanghai, China: IEEE, 2013, pp. 1–8.

[10] J. Kossaifi et al., "SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 1022–1040, Mar. 2021.

[11] K. Xie, V. W. Vongkulluksn, B. C. Heddy, and Z. Jiang, "Experience sampling methodology and technology: An approach for examining situational, longitudinal, and multi-dimensional characteristics of engagement," *Educ. Technol. Res. Develop.*, vol. 72, no. 5, pp. 2585–2615, 2024.

[12] M. Csikszentmihalyi, *Flow: The Psychology of Optimal Experience*. New York, NY, USA: Harper Perennial, 1990.

[13] T. Baltrušaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Piscataway, NJ, USA: IEEE, 2018, pp. 59–66.

[14] D. Rae Selvig and H. Schoenau-Fog, "Non-intrusive measurement of player engagement and emotions—Real-time deep neural network analysis of facial expressions during game play," in *HCI in Games*, X. Fang, Ed., Cham, Switzerland: Springer International Publishing, 2020, pp. 330–349.

[15] A. R. Clarke, R. J. Barry, R. McCarthy, M. Selikowitz, and S. J. Johnstone, "Effects of stimulant medications on the EEG of girls with attention-deficit/hyperactivity disorder," *Clin. Neurophysiol.*, vol. 118, no. 12, pp. 2700–2708, 2007.

[2] https://github.com/AmmarRashed/MultimodalEngagement