Real-Time Prediction of Player Engagement from Multimodal Data

Ammar Rashed, Shervin Shirmohammadi, *Fellow, IEEE,* Ihab Amer, *Senior Member, IEEE,* Mohamed Hefeeda *Senior Member, IEEE,*

***Abstract*—Predicting player engagement in video games offers significant advantages for game design and quality assessment. Current approaches for predicting engagement, however, are costly, invasive, and/or suffer from recall bias, which limits their employability in practice. This paper presents a new multimodal neural network model for engagement prediction and conducts systematic analyses to identify practical solutions for real-world applications. We develop a hybrid architecture combining convolutional downsampling with transformer-based sequence modeling that effectively processes temporal data from multiple sensing channels, including EEG, eye tracking, and webcam footage. Through comprehensive comparative analysis, we determine which modalities and specific features most strongly contribute to engagement detection. Our evaluation demonstrates that webcam-based approaches utilizing geometric facial features and deep learning-based emotional embeddings achieve performance comparable to specialized biometric equipment while requiring only standard hardware. Feature attribution analysis reveals that head pose dynamics and facial embeddings are the most influential engagement signals, with neurophysiological features showing limited contribution. Cross-modal analysis confirms that facial emotion-based combinations achieve the strongest performance. Statistical power analysis reveals participant variability, highlighting generalization challenges. These findings demonstrate that effective engagement measurement is achievable using accessible webcam technology, offering an optimal balance between accuracy and accessibility for widespread commercial implementation in gaming contexts.**

***Index Terms*—Player engagement, engagement prediction, gaming systems.**

I. Introduction

In video games, player engagement refers to the deep and multifaceted psychological and emotional connection players establish with a game. It encompasses the player's cognitive, affective, and behavioral states and interactions during gameplay. Player engagement has a central role in human-computer interaction [1], significantly influencing how players perceive and interact with video games. It extends beyond mere interaction to encompass the player's immersion, motivation, and overall satisfaction [2], [3]. As games increasingly embrace narrative elements, the assessment of gaming systems extends beyond technical parameters like latency and frame rate. Consequently, the concept of player engagement has garnered attention from both academia and industry, underscoring its importance in game design and optimization [4]. Engaged players, characterized by heightened concentration and enjoyment, contribute to the appeal of video games. Effective assessment of player engagement is essential for understanding player experiences [5] and offers valuable insights for the optimization of game quality, resource allocation in cloud gaming [6], [7], and user satisfaction [8].

Measuring player engagement, however, presents significant challenges. Traditional approaches rely on post-gameplay questionnaires that introduce recall bias [9], or employ biometric sensors that require specialized laboratory equipment, limiting their practical application in real-world settings. While game telemetry provides valuable behavioral insights, it often lacks generalizability across different games and may not directly capture the player's subjective experience. An ideal engagement measurement system should be non-interrupting, based on player-reported ground truth, minimize temporal bias, work across different games, and be deployable with commonly available hardware.

In this paper, we present multiple research contributions using the open-source MultiPENG dataset [10]. First, we develop and evaluate a hybrid neural architecture that effectively processes temporal data from multiple sensing channels, including EEG, eye tracking, and webcam footage. Second, we conduct a systematic comparative analysis to determine which modalities and specific features most strongly contribute to engagement detection. Third, we demonstrate that webcam-based approaches, which utilize only standard hardware available on most computers, can achieve comparable performance to specialized biometric equipment, offering an optimal balance between accuracy and accessibility for real-world applications.

While our analysis incorporates multiple sensing modalities (EEG, eye tracking, webcam) to establish performance benchmarks and understand which signals contribute to engagement detection, our primary contribution is demonstrating that webcam-based approaches alone achieve comparable performance (63% vs 65% accuracy) using only standard hardware, thereby eliminating the need for specialized equipment in practical deployment. Through detailed ablation studies, feature attribution analysis, and cross-modal comparisons, we provide insights into how different sensing modalities capture complementary aspects of player engagement. Our findings reveal that head pose dynamics and deep learning-based facial embeddings serve as particularly strong engagement indicators,

A. Rashed and S. Shirmohammadi are with the University of Ottawa, Ottawa, Ontario, Canada (e-mail: arasi005@uottawa.ca; shervin.s@uottawa.ca).

I. Amer is with American University of Sharjah (AUS), Sharjah, United Arab Emirates (e-mail: iamer@aus.edu).

M. Hefeeda is with Simon Fraser University, Burnaby, British Columbia, Canada (e-mail: mhefeeda@sfu.ca).

suggesting that webcam data alone can capture much of the engagement-relevant information without requiring specialized hardware. These results have significant implications for developers and researchers seeking to implement engagement-aware systems in gaming applications beyond controlled laboratory environments, with potential adaptations for educational contexts. Dataset and analysis ethically approved (UOttawa Ethics H-07-23-9439) with consent.

The remainder of this paper is organized as follows: Section II reviews background concepts and related work in engagement measurement. Section III details our data processing pipeline and model architecture. Section IV presents experimental results and comparative analyses across sensing modalities. Finally, Section V summarizes our findings and discusses implications for future engagement-aware applications.

II. Background and Related Work

A. Background

1) Definition of Engagement: Player engagement is a multidimensional construct that encompasses cognitive, affective, and behavioral aspects of the gameplay experience [11]. Several theoretical frameworks have been proposed to explain engagement. Flow theory describes engagement as the optimal balance between game challenge and player skill [12], with practical formulations including the measurement model [13] and experience fluctuation model [14]. Alternative perspectives include sensory and imaginative immersion [15], cognitive absorption [16], emotional attachment [17], and motivational aspects like conation [18]. Chen et al. define engagement as a "sustained level of involvement caused by capturing a person's interest, holding the majority of a person's attentional resources, and placing the person in an immersive state" [19].

2) Importance of Engagement: The importance of engagement extends beyond player experience to business metrics. Engagement directly correlates with monetization strategies, particularly in freemium games where playing time influences exposure to advertisements and in-app purchases [20]. Engagement metrics can predict player churn, activity patterns, and purchase propensity, enabling timely interventions for retention and monetization [21]. Different game genres exhibit distinct engagement dynamics; for instance, in multiplayer games, interactions with other players strongly influence engagement levels. Karmakar et al. demonstrated that jointly modeling players' motivation, progression, and churn in role-playing games significantly improved retention analysis [22].

Beyond commercial applications, engagement serves as a key metric for measuring the efficacy of educational games [23], entertainment games [24], and evaluating cognitive effects of gaming [25]. Strategic focus on engagement can enhance purchase intent, word-of-mouth marketing, and player recruitment [26]. A recent comprehensive review by Rashed et al. presents a taxonomy of engagement measurement approaches across different modalities, identifying critical trade-offs between complexity and interpretability, accuracy and generalizability [27].

B. Related Work

Current approaches to engagement measurement face several limitations that affect their practicality and reliability. We categorize and summarize the key approaches in the literature in the following.

1) Self-Report Methods: Standard engagement questionnaires include the Game Engagement Questionnaire [28], the Immersive Experience Questionnaire [29], and Player Experience of Need Satisfaction [30]. While these instruments capture various aspects of engagement, they are typically administered after gameplay, introducing recall bias [9] and failing to capture moment-to-moment fluctuations in engagement. Semi-structured interviews provide deeper insights into player experiences [31] but share the same temporal limitations, making both approaches unsuitable for real-time measurement.

While self-report measures capture subjective experiences that may be influenced by mood, fatigue, or other individual factors, these influences are inherent to the construct being measured rather than confounds to be eliminated. In engagement prediction, the objective is to predict what players will report, incorporating whatever factors affect their subjective experience.

2) Biometric Signals: Physiological and neurological measurements provide objective indicators of engagement. EEG signals have been used to develop engagement indices based on frequency bands [32], though signal flickering presents challenges for consistent measurement [33]. Apicella et al. demonstrated that EEG-based engagement detection systems can achieve reliable performance in interactive environments where both cognitive and motor skills are engaged, using filter bank and common spatial pattern features with support vector machines [34]. Eye-tracking metrics like pupil dilation, blink rate, and gaze direction offer insights into visual attention [35], with some metrics approximable from facial footage via webcams [36]. Additional physiological measures include heart rate, respiratory rate, and electrodermal activity [37]. While biometric data provide deeper insights, they require specialized equipment that limits studies to laboratory settings and raises privacy concerns.

3) Multimodal Approaches: Several studies have explored combining multiple data sources to measure engagement more comprehensively. Wu et al. demonstrated that combining video features with speech and audio modalities improves engagement detection performance, and introduced techniques to address challenges of label quality and intra-class variety in engagement measurement [38]. EngageMon [39] integrates touch events, physiological signals, and motion data for mobile gaming contexts. Pinitas et al. combined gameplay frames and gamepad inputs to predict engagement [40]. While promising, these approaches face challenges in generalizability across gaming contexts and effective modality integration.

4) Game-Specific Elements: Game telemetry provides insights into player behavior but often lacks generalizability across different games. Pinitas et al. used game scene frames and gamepad actions to estimate real-time engagement [40], though their method requires separate models for different games. Reguera et al. measured engagement based on level completion attempts and abandonment time [41], assuming

standardized difficulty levels across games. Rashed et al. developed a framework for non-intrusive, real-time engagement measurement using game telemetry data, focusing on online multiplayer games [42]. While telemetry-based approaches offer valuable insights, they still require game-specific adaptations, unlike biometric methods which function independently of game mechanics.

5) Webcam Video: Webcam data offers a non-intrusive, accessible source for measuring engagement through facial expressions, head movements, and gaze information. Previous works have used facial data to quantify emotions in educational [43] and gaming contexts [44]. FaceEngage used webcam feeds to measure engagement based on facial embeddings, head motion, gaze motion, and facial action units [24], though it relied on game status assumptions rather than player-reported ground truth. Rae et al. measured engagement using facial expressions with continuation desire (conation) as ground truth [18], though this approach employs a narrow definition of engagement focused primarily on motivation. The main limitation of existing webcam approaches is their reliance on narrow engagement definitions rather than comprehensive player-reported measures.

C. Requirements for Effective Engagement Measurement

Based on the limitations of existing approaches, we establish five key criteria that an effective player engagement measurement method should satisfy:

- **C1. Non-interrupting:** The method should not disrupt the natural flow of gameplay. Interruptions can introduce noise and alter the gaming experience, making continuous measurement unsustainable.
- **C2. Player-Reported Labels:** Indirect proxies capture limited aspects of engagement. The ground truth should derive from players' self-reported experiences rather than assumptions about gameplay status or narrow emotional indicators.
- **C3. Unbiased:** The temporal gap between experience and reporting should be minimized to reduce recall bias. Traditional post-gameplay surveys fail to capture the dynamic nature of engagement throughout a long session.
- **C4. Game Generic:** The approach should generalize across different game genres, mechanics, and platforms. Methods tied to specific game elements lack broader applicability.
- **C5. Practical:** For real-world deployment, the method should not require inaccessible equipment or necessitate sharing sensitive information.

As summarized in Table I, existing methods each fail to satisfy one or more of these criteria. Surveys and interviews [45], while gathering player-reported data, are inherently interruptive and vulnerable to recall bias. Biometric approaches [32], [37] provide non-interrupting, real-time measurements across game contexts but require specialized equipment and access to sensitive data. Game-specific methods [40]–[42] operate non-intrusively but require adaptations across different games and often rely on proxy indicators rather than player reports. Standard webcam methods [18], [24] offer practical, non-interrupting measurement across game types but typically rely on narrow engagement definitions rather than comprehensive player-reported ground truth.

TABLE I
CRITERIA COMPARISON OF EXISTING METHODS

Method	C1	C2	C3	C4	C5
Surveys & Interviews	✗	✓	✗	✓	✓
Biometric Signals	✓	✓	✓	✓	✗
Game-Specific Elements	✓	✗	✓	✗	✓
Standard Webcam Methods	✓	✗	✓	✓	✓
Proposed method (MultiPENG Webcam)	✓	✓	✓	✓	✓

To address these limitations, our work focuses on the webcam-based modality within the MultiPENG dataset [10]. Unlike previous webcam approaches such as FaceEngage [24] that infer engagement from game status, or conation-focused methods [18], the webcam data in MultiPENG is aligned with authentic player-reported ground truth. The dataset collection employs strategic timing of self-report measures during natural gameplay pauses, reducing the duration of annotated sessions to an average of 46 seconds for a minimal recall bias. This approach enables the development of webcam-based models that satisfy all five criteria: they are non-interrupting, based on player-reported ground truth, minimize bias, work across different games, and only require a standard webcam.

Regarding privacy considerations for webcam-based approaches, we note that webcams are already ubiquitous in modern gaming environments, particularly among streamers and players engaged in multiplayer games with video chat features. Unlike biometric sensors that capture physiological data, webcam processing for engagement detection can be performed locally on the user's device without transmitting video data, with only engagement metrics shared if desired. This processing model aligns with existing privacy expectations in gaming contexts where players routinely use webcams for streaming and communication purposes.

A key advantage of using the MultiPENG dataset is that it also provides aligned multimodal data, including EEG and eye-tracking information synchronized with webcam footage. This allows us to develop multimodal baselines against which we can validate the performance of our webcam-based approach. The multimodal nature of the dataset enables comparative analysis across different sensing modalities while maintaining consistent ground truth measures.

In the following sections, we describe our methodology for utilizing the webcam modality from this dataset to build a robust, game-generic engagement measurement system that satisfies all five criteria in Table I.

III. PROPOSED SOLUTION

A. Proposed Engagement Prediction Model

Before detailing our model architecture, we clarify the operational workflow. Our system is trained using synchronized multimodal data paired with player-reported engagement labels collected via ESM during natural gameplay pauses.

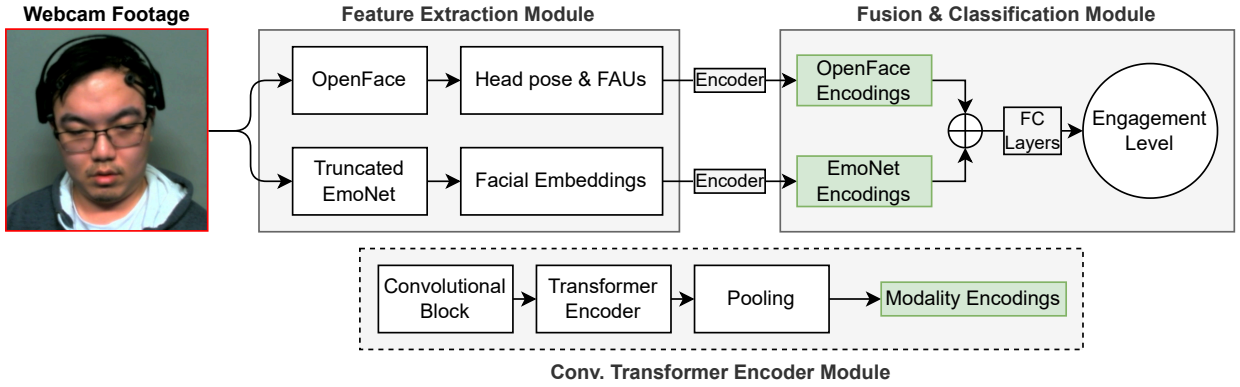


Fig. 1. Proposed multimodal model: each modality is encoded by a separate Convolutional Transformer before late fusion.

Once trained, the model can predict engagement from sensor data alone during deployment, eliminating the need for questionnaires during actual gameplay sessions. This enables real-time engagement assessment for game designers and developers based on models grounded in authentic player-reported experiences.

We design a neural network model to capture temporal engagement dynamics while handling the high dimensionality and varied sampling rates of our multimodal data. Our model balances complexity against dataset size constraints to prevent overfitting while enabling effective cross-modal feature extraction.

The model combines convolutional downsampling, transformer-based sequence modeling, and global pooling to obtain fixed-length feature vectors per modality, as shown in Figure 1. This design efficiently processes variable-length sequential data, maintains modality-specific processing without information loss, and captures both local patterns and long-range dependencies.

The model comprises two main components. The first is a Convolutional Transformer Encoder, which processes each modality independently through a 1D convolutional layer (kernel size 3, stride 2) that reduces sequence length while projecting input features to a uniform hidden dimension. A lightweight transformer (1-2 layers, single attention head) then captures dependencies between timesteps. Finally, an adaptive average pooling layer produces fixed-length representations regardless of input sequence length.

The second component is for multimodal fusion. We implement late fusion by processing modalities through separate encoder instances, concatenating the resulting feature vectors, projecting to a unified hidden representation via a fully-connected layer, and performing binary classification through a sigmoid activation. We chose late fusion because it allows optimal processing of each modality according to its characteristics before integration, which is particularly important given the diverse nature of our data. Specifically, we treated the facial features and the emotional features as separate modalities to allow specialized processing before fusion.

B. Multimodal Dataset Processing

We utilize the open-source MultiPENG dataset [10], a comprehensive multimodal collection of player engagement data captured during gameplay sessions. The dataset contains synchronized recordings from 39 participants playing FIFA'23 and Street Fighter V at varying difficulty levels, with data collected across multiple sensing modalities: EEG signals from a 14-channel EMOTIV EPOC X headset (128 Hz), eye tracking data from a Gazepoint GP3 tracker (60 Hz), and facial expressions and head pose captured via webcam (30 Hz). A notable contribution of this dataset is its innovative approach to gathering engagement ground truth through an Experience Sampling Method (ESM) [46] administered during natural gameplay pauses (e.g., between FIFA matches, after Street Fighter rounds). This approach minimizes both gameplay disruption and recall bias compared to post-hoc questionnaires or continuous self-annotation methods. Participants provided self-reported engagement ratings on a 5-point Likert scale (0: Very Bored to 4: Very Engaged) after gameplay segments, resulting in 900 labeled sessions with an average engagement score of 2.87. We binarize the classification tasks following previous works [10], [24] by formulating the classification as lower or higher than average, where engagement levels 0-2 are labeled as "Low" (30% of samples) and levels 3-4 are labeled as "High" (70% of samples). The dataset employs participant-based cross-validation to ensure complete separation between training and test sets, preventing data leakage while maintaining the natural class imbalance that reflects typical gaming engagement distributions.

Our approach inherently accounts for individual differences in baseline mood, gaming preferences, environmental factors, and other personal characteristics through the self-reported ground truth methodology. Critically, the ground truth represents what players actually report experiencing, incorporating all factors that influence their subjective state. Since our objective is to predict self-reported engagement rather than infer an objective "true" engagement state independent of mood, factors affecting mood are part of the phenomenon being measured, not confounds to be controlled.

Furthermore, our experimental design includes multiple normalization strategies: (1) participants experience varying difficulty levels, creating an internal reference frame for comparing

their own engagement across sessions; (2) a warmup period allows players to calibrate their understanding of engagement in the gaming context [10]; and (3) binary classification uses the population-level average engagement (2.87) as the threshold, naturally accounting for individual differences in reporting tendencies. This approach prioritizes ecological validity and real-world applicability, as our goal is to develop systems that predict player-reported states across diverse individuals and natural gaming contexts.

We implemented modality-specific processing pipelines that extract relevant features while preserving temporal structure. Each modality was processed with optimized sampling rates determined through empirical testing of native rate, every second sample, and every third sample approaches. EEG, eye tracking, and OpenFace features were downsampled to every third sample, while EmoNet embeddings were processed at native rate, harmonizing temporal frequencies across modalities for optimal performance.

For EEG data, we utilized frequency band powers—theta (4-8 Hz), alpha (8-12 Hz), low beta (12-16 Hz), high beta (16-25 Hz), and gamma (25-45 Hz) bands—from all 14 channels as computed by the EMOTIV commercial software suite. These frequency bands correspond to different aspects of brain activity, with theta and alpha bands commonly associated with attention and relaxation states. The EMOTIV system provides continuous signal quality monitoring through an overall quality score (EQ.OVERALL) ranging from 0-100, which aggregates contact quality, signal quality, and signal magnitude metrics. Low-quality EEG signals were filtered by nullifying band power values when this overall quality score fell below 50%, with interpolation maintaining sequence continuity. This threshold was selected based on empirical analysis showing that only 50% of samples achieve an average EEG quality of at least 75%, primarily due to movement artifacts during gameplay that cause temporary signal degradation. We validate this threshold in Section IV-A1.

For eye tracking, we extracted fixation data (position and duration), saccade characteristics (magnitude and direction), pupil diameter, and blink patterns from the Gazepoint software. Missing values were interpolated before downsampling the 60Hz data.

Facial features were extracted using two methods: OpenFace [47] provided head pose dynamics (3D translation/rotation vectors) and facial action unit intensities (0-5) for all 17 detected units, with derivative features (velocity and acceleration) calculated between consecutive frames. We also extracted facial embeddings using EmoNet [48] by truncating the output layers.

We applied min-max scaling across all modalities to normalize feature values, with scaling parameters derived exclusively from the training set to prevent data leakage. This standardized pipeline maintains the unique temporal characteristics of each modality while ensuring reproducibility.

C. Model Training

We trained the model using an Adam optimizer with learning rate 5e-5 and Binary Cross-Entropy loss with class weighting to address imbalance. Training used a batch size of 64 samples for up to 150 epochs with early stopping (patience: 10 epochs). Our cross-validation ensured complete separation between participants in training, validation, and test sets.

To address our limited sample size (approximately 580 training samples), we employed regularization through dropout (0.3-0.5), early stopping based on validation performance, a relatively small model size (hidden dimension of 128, single attention head), and convolutional downsampling to reduce effective sequence length.

Different modalities in our dataset have varying sampling rates: webcam footage at 30Hz, EEG band data at 8Hz, and eye tracking at 64Hz. Rather than resampling to a common rate, we processed each modality at its native sampling rate to preserve fine-grained temporal information. We empirically validated this decision by comparing against models trained on resampled data (at 10Hz and 20Hz), finding that native-rate processing consistently outperformed resampled approaches.

Our participant-based cross-validation scheme used consistent inner splits. For each outer fold (test set), we used a single fixed train/validation split rather than multiple inner fold combinations, ensuring each test participant is evaluated once with a model trained on a consistent subset. This approach establishes a standardized evaluation framework enabling fair comparison while maintaining participant-based separation between sets.

The combination of convolutional and transformer-based processing allows the model to handle long sequences without requiring fixed-length inputs and with reduced susceptibility to overfitting. This architecture represents a balance between complexity and practicality, allowing direct comparison between sensing modalities by using identical architectures with modality-specific parameters.

D. Deployment Considerations

Our architectural design supports flexible deployment across different computational budgets and application scenarios. The lightweight architecture—single-head transformer with hidden dimension 128 and convolutional downsampling—balances predictive performance against computational efficiency for practical implementation on consumer gaming hardware.

To validate practical feasibility, we conducted runtime performance analysis on 100 representative samples from the dataset using consumer-grade hardware (Intel i9-10900K CPU, NVIDIA RTX 3080 GPU, 32GB RAM at 3200MHz). Feature extraction using OpenFace (CPU-only processing with OpenCV [49] and dlib [50]) requires 0.344 seconds per second of video (standard deviation (SD) = 0.008), with memory sampled at 100ms intervals averaging 856.2 MB (SD = 17.9) and peaking at 883.9 MB (SD = 25.7). Neural network inference for the webcam modality requires only 0.037 milliseconds per second of video (SD = 0.014), with minimal VRAM requirements (0.5 MB model size, 14.9 MB peak). The low coefficient of variation for feature extraction (2.27%) indicates consistent performance, while higher inference variance (39.18%) reflects variable-length sequence processing.

These measurements demonstrate real-time processing capability with substantial headroom—feature extraction operates

TABLE II
ENGAGEMENT PREDICTION PERFORMANCE ACROSS MODALITIES AND APPROACHES

Approach	F1-Score			ROC-AUC	Accuracy
	Low Engagement	High Engagement	Average		
Dummy (All High)	-	0.82 ± 0.03	0.41 ± 01	0.50	0.70 ± 0.04
Webcam (OpenFace + EmoNet)	0.35 ± 0.08	0.71 ± 0.04	0.53 ± 0.03	0.56 ± 0.04	0.63 ± 0.04
Eye Tracking	0.22 ± 0.09	0.63 ± 0.09	0.43 ± 0.02	0.48 ± 0.04	0.59 ± 0.06
EEG	0.31 ± 0.09	0.65 ± 0.07	0.48 ± 0.02	0.57 ± 0.06	0.61 ± 0.05
Multimodal (All)	0.38 ± 0.07	0.73 ± 0.04	0.56 ± 0.03	0.59 ± 0.04	0.65 ± 0.03

at approximately 3x real-time speed while model inference adds negligible overhead. The framework accommodates variable prediction frequencies: periodic assessment (every 10-20 seconds), event-driven prediction at natural game pauses, or batch analysis. The webcam-based approach operates on standard gaming hardware, while the multimodal variant can leverage GPU acceleration. Deployment can be client-side for privacy or server-side for computational offloading.

IV. EVALUATION

This section presents our engagement estimation model evaluation and key findings. We first compare performance across individual modalities and multimodal combinations (Section IV-A), revealing that webcam-based approaches provide an optimal balance of accuracy and accessibility. We then benchmark against human annotators (Section IV-C), demonstrating that our multimodal model outperforms human judges in classifying engagement states. Statistical power analysis (Section IV-D) identifies significant participant-specific variations that impact model performance. Through ablation studies (Section IV-E) and feature attribution analysis (Section IV-F), we establish that facial expressions and head movements contribute most significantly to engagement detection, with EmoNet embeddings providing the strongest signals. Our cross-modal analysis shows that combining EmoNet with either EEG or eye tracking yields the strongest pairwise performance. Finally, architectural analysis validates the effectiveness of our hybrid ConvTransformer design, particularly for processing heterogeneous multimodal data. All reported uncertainties represent the standard error of the mean (SEM) across cross-validation folds.

A. Player Engagement Prediction Results

We evaluated performance across individual sensing modalities and multimodal fusion, using standard classification metrics to assess effectiveness. We report F1-scores (harmonic mean of precision and recall) for both engagement classes and their average, ROC-AUC (area under the receiver operating characteristic curve, measuring discrimination ability), and overall accuracy. As shown in Table II, the performance varies across modalities, with all approaches demonstrating higher F1-scores for high engagement detection compared to low engagement. The multimodal approach achieved the highest ROC-AUC (0.59 ± 0.04). We analyze these patterns in detail below.

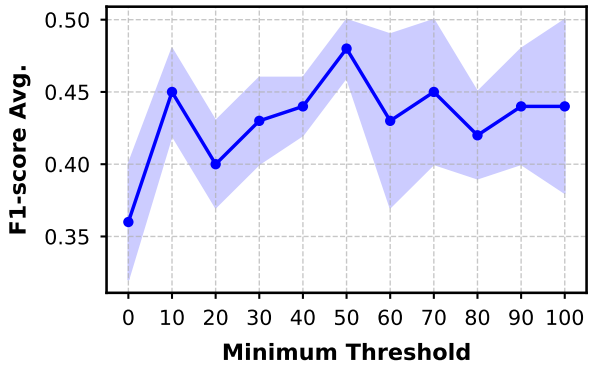


Fig. 2. Effect of EEG quality threshold on F1-score. Shaded area shows SEM.

1) Signal Quality Optimization: For EEG data, which is susceptible to noise and motion artifacts during gameplay, we conducted a systematic analysis to determine the optimal signal quality threshold. As shown in Figure 2, prediction performance improved with increasing quality thresholds up to approximately 50%, after which performance stabilized and eventually declined slightly at higher thresholds. This pattern indicates that excluding signals below the 50% quality threshold effectively removes noise while retaining informative data. We adopted this threshold for all EEG-based prediction results presented in Table II.

2) Unimodal Performance Analysis: Examining individual modalities in Table II, the webcam-based approach achieved the highest accuracy (63% ± 0.04) among single-modality approaches, with strong performance in identifying high engagement states (F1-score: 0.71 ± 0.04). The precision and recall for high engagement detection were 0.76 ± 0.03 and 0.69 ± 0.08, respectively (not shown in table). These results suggest that facial expressions and head movements provide valuable engagement indicators while offering significant practical advantages through standard hardware accessibility.

EEG features produced the second-highest unimodal accuracy (61% ± 0.05) with an average F1-score of 0.48 ± 0.02. While EEG showed the highest precision for detecting high engagement (0.78 ± 0.05), it demonstrated lower recall (0.64 ± 0.12). Despite this strong performance after quality filtering, the specialized equipment and calibration requirements present substantial implementation barriers in everyday gaming scenarios.

Eye tracking features showed the lowest unimodal performance (59% ± 0.06 accuracy), particularly struggling with low

TABLE III
ARCHITECTURE PERFORMANCE COMPARISON (AVERAGE F1-SCORE)

Approach	Architecture	F1 (Avg)	ROC-AUC
Multimodal	Hybrid	**0.56 ± 0.03**	0.59 ± 0.04
	Transformer-only	0.51 ± 0.03	0.52 ± 0.06
	Convolution-only	0.52 ± 0.03	0.59 ± 0.04
Webcam	Hybrid	0.53 ± 0.03	0.56 ± 0.04
	Transformer-only	0.49 ± 0.01	0.56 ± 0.03
	Convolution-only	**0.54 ± 0.03**	0.58 ± 0.04

engagement detection (F1-score: 0.22 ± 0.09).

3) Multimodal Integration Benefits: Our multimodal approach combining all sensing channels achieved the highest overall performance among our models, with 65% accuracy (± 0.03) and an average F1-score of 0.56 ± 0.03, as shown in Table II. This represents modest but consistent improvements across all metrics compared to the best single-modality approach. The fusion model showed particular gains in low engagement detection (F1-score: 0.38 ± 0.07 compared to 0.35 ± 0.08 for webcam alone), suggesting that different modalities capture complementary aspects of engagement.

However, the practical significance of this 2 percentage point improvement in accuracy over the webcam-only approach must be weighed against the substantial increase in implementation complexity, which requires synchronization of multiple specialized sensors and more computational resources.

B. Architecture Analysis

To validate our hybrid ConvTransformerEncoder design, we compared it against transformer-only and convolution-only variants. Each variant maintained identical hidden dimensions, training configurations, and input/output interfaces.

Results in Table III reveal distinct patterns across sensing modalities. For multimodal data, the hybrid architecture achieved the highest F1-score (0.56 ± 0.03), demonstrating that combining convolutional feature extraction with transformer sequence modeling provides complementary benefits for heterogeneous sensing data. The convolution-only architecture showed slight advantages for webcam data (0.54 ± 0.03), though this difference falls within statistical error margins.

The transformer-only architecture consistently demonstrated poor class balance, with low-engagement F1-scores as low as 0.22 ± 0.07 for webcam data despite reasonable overall accuracy. This imbalance makes it unsuitable for applications requiring balanced engagement detection.

These findings support our hybrid architecture as the most robust approach across different sensing modalities. While specialized architectures might offer marginal improvements in specific scenarios, the hybrid ConvTransformer provides consistent performance across input types and engagement states—critical for real-world deployment where contexts vary. The architecture effectively leverages convolutional processing for local pattern extraction and transformer mechanisms for capturing temporal dependencies, aligning with established approaches in physiological signal processing [51].

C. Human Annotation Analysis

To establish comparative benchmarks, we evaluated our machine learning models on the same 20 samples that human annotators rated. These samples came from 5 participants whose data were completely excluded from training and validation sets to prevent data leakage. We applied both individual fold-specific models (reporting average performance) and an ensemble approach (averaging predictions from all 7 models). Two approaches were compared: the webcam-based model (processing the same visual information available to human judges) and the multimodal approach (incorporating all sensing channels).

As shown in Table IV, several notable patterns emerge. The multimodal ensemble approach achieved the highest performance (56% accuracy, 0.63 ROC-AUC), modestly outperforming human annotators (50% accuracy). When comparing ensemble to non-ensemble results, interesting trends emerge: while the multimodal approach benefited from ensemble averaging (56% vs. 52% accuracy), the webcam-based model's performance decreased (40% vs. 46% accuracy). This divergent pattern suggests that ensemble averaging helps stabilize predictions when incorporating physiological signals but may amplify biases when relying solely on visual features for this challenging subset.

The webcam-based model underperformed compared to human annotators on this challenging subset (40-46% accuracy vs. 50%), despite using the same visual information. This performance, however, falls below the same model's capabilities on the full dataset (63% accuracy), suggesting these particular samples present specific prediction challenges. The performance degradation was more pronounced in the ensemble approach for the webcam model, indicating that averaging predictions across multiple models may reinforce rather than mitigate biases when the underlying features lack robust engagement signals.

It's important to contextualize these findings with the previously reported low inter-rater agreement among human annotators in (Krippendorff's $\alpha = 0.040$) [10], highlighting engagement's nature as a complex internal experience rather than an easily observable state.

The multimodal approach demonstrated improved discrimination ability, particularly for low engagement states (precision of 0.83 for the ensemble model compared to 0.59 for human annotators). This suggests that augmenting webcam-based features with eye tracking and EEG data enables the model to capture aspects of engagement not visually apparent to human judges.

Examining the class distribution in model predictions reveals important patterns: the webcam model favors high engagement detection (recall of 0.75-0.77 for high vs. 0.17-0.26 for low engagement), while the multimodal approach achieves more balanced detection. This suggests that physiological signals provide critical information for identifying low engagement states that may not manifest through observable facial expressions or behaviors alone.

The performance gap between the full dataset results (63% accuracy for webcam, 65% for multimodal) and this challenging subset raises important questions about participant-specific

TABLE IV
DETAILED PERFORMANCE COMPARISON ON HUMAN ANNOTATION SUBSET (20 SAMPLES)

Approach	Variant	F1-Score			ROC-AUC	Accuracy
		Low Engagement	High Engagement	Average		
Human Annotators	—	0.55 ± 0.04	0.40 ± 0.04	0.47 ± 0.03	—	0.50 ± 0.03
Webcam	No Ensemble	0.32 ± 0.09	0.53 ± 0.02	0.43 ± 0.05	0.50 ± 0.03	0.46 ± 0.04
	Ensemble	0.25	0.50	0.38	0.46	0.40
Multimodal	No Ensemble	0.44 ± 0.12	0.52 ± 0.02	0.48 ± 0.06	0.61 ± 0.04	0.52 ± 0.05
	Ensemble	**0.56**	**0.56**	**0.56**	**0.63**	**0.56**

sensitivity in engagement modeling. The webcam-based approach showed a more substantial performance drop on these specific samples compared to the multimodal approach, suggesting higher susceptibility to individual differences. This pattern resembles the ”cold-start problem” commonly observed in personalized modeling systems, which we investigate further through statistical power analysis in the following section.

It is important to contextualize these F1-score metrics within the broader challenge of measuring subjective psychological states. F1-scores represent the harmonic mean of precision and recall, indicating how reliably the system identifies both high and low engagement states compared to player self-reports—our most direct measure of subjective experience. Our models do not claim to predict players' ”exact” internal experiences; such precision is neither achievable nor necessary for practical applications. Rather, the multimodal model's average F1-score of 0.56 (compared to 0.47 for human annotators) demonstrates that the system provides consistent, balanced detection of engagement states despite the class imbalance in our dataset. For game designers and stakeholders, this level of performance—particularly the balanced F1-scores across both engagement classes—enables data-driven decisions about game design iterations, difficulty balancing, and user experience optimization. These applications benefit from reliable directional insights about engagement patterns rather than requiring perfect prediction of individual internal states.

D. Statistical Power Analysis

To understand the performance discrepancy between the full dataset and the human annotation subset, we conducted a statistical power analysis examining whether our dataset contained sufficient samples to detect meaningful differences.

We calculated Cohen's d effect sizes and statistical power using the non-central t-distribution approach for F1-score and ROC-AUC metrics. Analysis of the webcam-based model revealed substantial fluctuations in effect sizes across different sample sizes, ranging from d=0.65 (full dataset) to d=0.96 (n=5). The multimodal approach showed even greater variability, with effect sizes ranging from d=0.64 to d=2.27.

As shown in Figure 3, even with modest effect sizes, the webcam-based model requires approximately 20-25 samples to achieve the conventional 80% power threshold. These fluctuations highlight a phenomenon similar to the “cold-start problem” in personalized modeling systems [52]—model performance instability when incorporating data from new

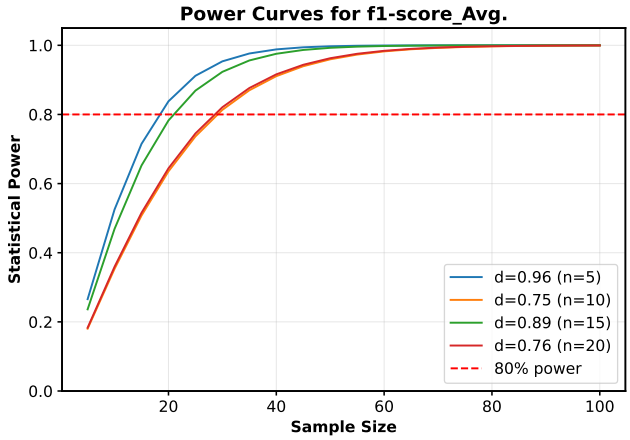


Fig. 3. Statistical power curves for F1-score averages of the webcam-based model. Dashed line marks the 80% power threshold.

participants. This sensitivity is particularly pronounced for webcam-based features, which rely on visual cues that vary significantly across individuals.

For the full dataset, F1-score power reached 0.9976, exceeding the conventional threshold, while ROC-AUC power reached 0.7456, slightly below optimal. This difference suggests ROC-AUC exhibits higher variance across cross-validation folds, requiring larger sample sizes for comparable power.

These results demonstrate that while our sample size is statistically robust for primary analyses, model performance—particularly for the webcam-based approach—remains sensitive to specific participant composition. This sensitivity explains the performance variability between the full dataset and specific subsets, highlighting inherent challenges in modeling subjective psychological states across diverse individuals.

E. Ablation Study

To quantify each sensing channel's contribution within our multimodal approach, we conducted an ablation study by systematically removing individual modalities from the full model.

We established a baseline using the complete model with all four modalities (EEG, eye tracking, OpenFace, and EmoNet) across 7 cross-validation folds. We then evaluated four variant

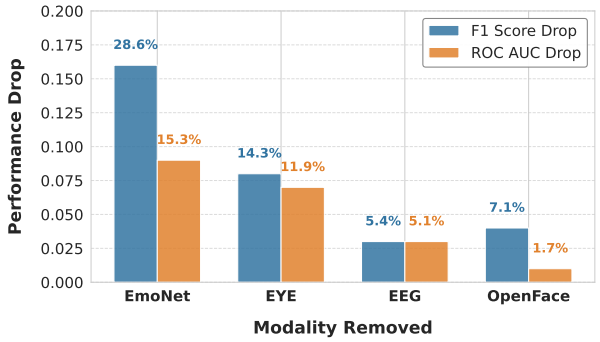


Fig. 4. Performance drop from removing each modality, relative to the full multimodal model.

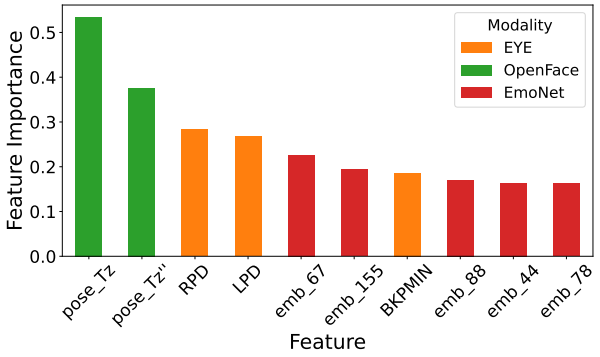


Fig. 5. Top 10 features ranked by absolute attribution importance.

models, each with one modality removed, measuring performance drops in F1 score and ROC AUC.

As shown in Figure 4, the results reveal a clear hierarchy of modality importance. EmoNet facial embeddings provide the most substantial contribution, with removal causing a 28.6% reduction in F1 score and 15.3% drop in ROC AUC. Eye tracking emerged as the second most influential modality (14.3% decrease in F1 score, 11.9% in ROC AUC). OpenFace features and EEG signals showed more modest contributions (7.1% and 5.4% drops in F1 score, respectively).

These findings reinforce that webcam-derived features offer optimal balance between performance and accessibility, with EEG providing minimal additional benefit despite substantial implementation complexity.

F. Feature Attribution Analysis

To identify which specific signals most influence engagement classification, we conducted feature attribution analysis using integrated gradients [53] via Captum [54]. This analysis quantifies each feature's contribution toward predicting high engagement (positive attributions) versus low engagement (negative attributions).

Figure 5 reveals that facial and eye tracking features dominate the importance ranking. Head pose translation along the z-axis (`pose_Tz`)—representing forward-backward movement—emerged as the most influential feature, followed by its acceleration. This suggests engaged players maintain consistent forward orientation, while movements away from

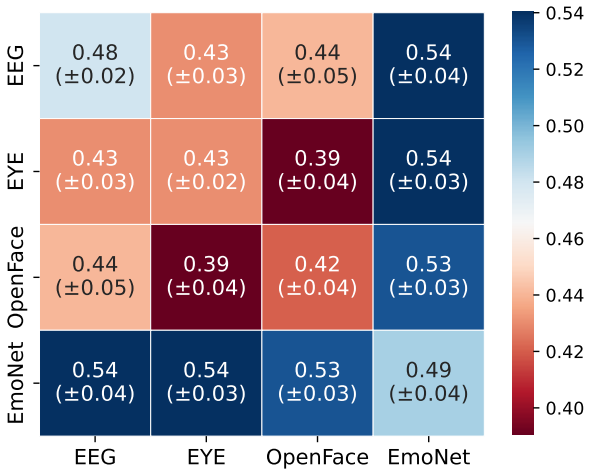


Fig. 6. F1-score heatmap for pairwise modality combinations. Cells show mean ± SEM; darker blue indicates better performance.

the screen indicate disengagement. Pupil dilation and blink rate also ranked highly, consistent with literature linking pupil responses to cognitive engagement [35].

EmoNet facial embeddings comprised 7 of the top 10 features for high engagement prediction and showed strong bidirectional attribution, highlighting their capacity to capture nuanced emotional states beyond geometric features. Among EEG features, only frontal theta activity (AF3 and F8 electrodes) appeared in top rankings, aligning with established research on theta oscillations as cognitive engagement markers [32], [55].

Features showing strong attribution to both engagement states reveal non-linear relationships between physiological signals and engagement, explaining why multimodal approaches demonstrated more balanced detection capabilities compared to single-modality methods. The dominance of facial and eye tracking features confirms that accessible webcam-derived features capture the most engagement-relevant information, strongly supporting practical implementation of webcam-based engagement measurement systems.

G. Cross-Modal Analysis

While our ablation study identified the contribution of each modality to the full model, we now examine how modalities complement each other when paired. This cross-modal analysis quantifies how different modality combinations contribute to engagement detection and identifies which pairs offer the strongest synergistic benefits.

We trained binary engagement classifiers for each possible pair of modalities (EEG, Eye, OpenFace, and EmoNet) using our ConvTransformerEncoder architecture with late fusion, maintaining the same training configuration and participant-based cross-validation approach described in Section III.

Figure 6 reveals important patterns in modality complementarity. Combinations involving EmoNet generally achieve the strongest performance, with EmoNet+EEG and EmoNet+Eye pairs both achieving the highest F1-score of 0.54±0.04 and 0.54±0.03 respectively. This suggests that facial emotional

expressions captured by EmoNet provide complementary information to both neurological and visual attention signals. Conversely, the OpenFace+Eye combination showed the weakest performance (0.39±0.04), indicating potential redundancy between structural facial features and eye movements.

ROC-AUC analysis shows similar patterns with some notable differences. The EmoNet+EEG combination achieves the highest ROC-AUC score (0.59±0.04), indicating superior discrimination capability across different thresholds. EmoNet+OpenFace and EmoNet+Eye combinations also perform strongly (0.58±0.03/0.04), while eye tracking alone shows the lowest ROC-AUC (0.48±0.04), confirming limitations in using eye-tracking data in isolation.

These findings provide important insights for practical multimodal engagement detection systems and complement our ablation study by quantifying how pairs of modalities perform together rather than measuring performance drops from removing individual components. EmoNet's consistently strong performance across multiple combinations confirms the value of deep learning-based facial emotion embeddings for engagement detection. The strong performance of EmoNet+EEG suggests neurological signals capture engagement aspects not apparent in facial expressions, potentially reflecting internal cognitive states without visible manifestations. Meanwhile, the relatively poor performance of OpenFace+Eye indicates redundancy between these modalities, suggesting both capture similar aspects of visual attention and facial behavior. Given EmoNet-based combinations' superior performance compared to the more specialized hardware required for EEG and eye tracking, these results support prioritizing advanced webcam processing for practical applications.

H. Summary

Our evaluation reveals three principal insights: (1) Webcam-based approaches using facial embeddings and head pose achieve 63% accuracy, approaching the 65% multimodal performance while requiring only standard hardware; (2) The multimodal ensemble outperforms human annotators (56% vs. 50% accuracy) on challenging samples; (3) Participant-specific variations create a "cold-start problem," with model performance sensitive to individual differences in engagement expression. These findings establish that accessible webcam technology captures sufficient engagement signals for practical deployment.

V. Conclusion

We presented a multimodal neural network model for player engagement prediction that bridges the gap between practicality and performance. Player engagement prediction remains a challenging problem, as evidenced by the low inter-rater agreement among human annotators [10] and the limitations of existing approaches that suffer from recall bias, require specialized equipment, or rely on game-specific behavioral proxies.

Our hybrid architecture achieved 65% accuracy when combining EEG, eye tracking, and webcam data, while the webcam-only approach reached 63% accuracy using just standard hardware. On challenging samples where human annotators achieved 50% accuracy, our webcam approach performed within comparable range, while the multimodal ensemble surpassed human judges at 56%. These results demonstrate that accessible sensing approaches can match both specialized equipment and human judgment.

Through systematic analyses, we identified that head pose dynamics and deep learning-based facial embeddings provide the strongest engagement signals, with neurophysiological features showing limited additional contribution. Cross-modal analysis confirmed that facial emotion-based combinations achieve the highest performance, validating that standard webcams capture the majority of engagement-relevant information. However, statistical power analysis revealed substantial participant-specific variations, highlighting the "cold-start problem" where model performance fluctuates with new players.

These findings eliminate the primary deployment barrier for engagement-aware systems. Rather than requiring specialized biometric equipment, developers can implement engagement measurement using hardware already present in most gaming setups, enabling new possibilities for adaptive gaming experiences and quality assessment tools in natural environments.

Despite these contributions, our binary classification approach simplifies the nuanced engagement spectrum, and participant sensitivity suggests generalization challenges across diverse populations. Future work should explore personalized calibration techniques, continuous engagement measures, and domain adaptation methods. Integrating real-time game telemetry with webcam-based measurements offers promising avenues for creating responsive systems that dynamically adapt to individual player states.

While our work demonstrates effective engagement measurement using accessible webcam technology in entertainment gaming contexts, several considerations merit attention for broader deployment. First, although our models were developed and validated on competitive gaming scenarios (FIFA 23 and Street Fighter V), adaptation to educational applications would require careful calibration and validation, as task structures and engagement manifestations may differ across domains. Second, head pose dynamics—our most influential feature—may exhibit different patterns depending on display characteristics (screen size, viewing distance) and input modalities (controller versus mouse/keyboard), though the underlying engagement-related movements (forward lean indicating engagement, backward lean indicating disengagement) likely generalize across contexts. While our webcam-based approach provides a practical foundation for engagement measurement in gaming, domain-specific validation is recommended before deployment in substantially different contexts.

In summary, our work demonstrates that effective engagement measurement is achievable using readily available webcam technology. By establishing performance benchmarks and identifying key engagement indicators, we provide a foundation for developing personalized gaming experiences without specialized equipment or intrusive measurement techniques.

REFERENCES

[1] H. L. O'Brien, I. Roll, A. Kampen, and N. Davoudi, "Rethinking (dis) engagement in human-computer interaction," *Computers in human behavior*, vol. 128, p. 107109, 2022.

[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," *Proceedings of the ACM on Human-Computer Interaction*, 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," *Behaviour & information technology*, pp. 1–17, 2023.

[4] L. A. Gil-Aciron, "The gamer psychology: a psychological perspective on game design and gamification," *Interactive Learning Environments*, vol. 32, no. 1, pp. 183–207, 2024.

[5] B. C. Strubberg, T. J. Elliott, E. P. Pumroy, and A. E. Shaffer, "Measuring fun: A case study in adapting to the evolving metrics of player experience," *Loading*, vol. 13, no. 21, pp. 1–19, 2020.

[6] A. A. Laghari, H. He, K. A. Memon, R. A. Laghari, I. A. Halepoto, and A. Khan, "Quality of experience (qoe) in cloud gaming models: A review," *multiagent and grid systems*, vol. 15, no. 3, pp. 289–304, 2019.

[7] I. Slivar, L. Skorin-Kapov, and M. Suznjevic, "Qoe-aware resource allocation for multiple cloud gaming users sharing a bottleneck link," in *2019 22nd conference on innovation in clouds, internet and networks and workshops (ICIN)*. IEEE, 2019, pp. 118–123.

[8] K. Xiaohan, M. N. A. Khalid, and H. Iida, "Player satisfaction model and its implication to cultural change," *IEEE Access*, vol. 8, pp. 184 375–184 382, 2020.

[9] E. Hassan, "Recall bias can be a threat to retrospective and prospective research designs," *The Internet Journal of Epidemiology*, vol. 3, no. 2, pp. 339–412, 2006.

[10] A. Rashed, S. Shirmohammadi, and M. Hefeeda, "Descriptor: Multimodal dataset for player engagement analysis in video games (multipeng)," *IEEE Data Descriptions*, vol. 2, pp. 17–25, 2025.

[11] W. Yang, M. Rifqi, C. Marsala, and A. Pinna, "Towards better understanding of player's game experience," in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, ser. ICMR '18. NY, USA: ACM, 2018, pp. 442–449.

[12] B. Cowley, D. Charles, M. Black, and R. Hickey, "Toward an understanding of flow in video games," *Computers in Entertainment (CIE)*, vol. 6, no. 2, pp. 1–27, 2008.

[13] G. B. Moneta, "On the measurement and conceptualization of flow," *Advances in flow research*, pp. 23–50, 2012.

[14] M. Bassi and A. Delle Fave, *Flow in the Context of Daily Experience Fluctuation*. Cham: Springer, 2016, pp. 181–196.

[15] L. Ermi and F. Mäyrä, "Fundamental components of the gameplay experience: Analyzing immersion," *Worlds in play: International perspectives on digital games research*, vol. 21, p. 37, 2007.

[16] B. Patzer, B. Chaparro, and J. R. Keebler, "Developing a model of video game play: motivations, satisfactions, and continuance intentions," *Simulation & Gaming*, vol. 51, no. 3, pp. 287–309, 2020.

[17] R. Leroy, "Immersion, flow and usability in video games," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–7.

[18] 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: Springer, 2020, pp. 330–349.

[19] M. Chen, B. E. Kolko, E. Cuddihy, and E. Medina, "Modeling but not measuring engagement in computer games," in *Proceedings of the 7th International Conference on Games + Learning + Society Conference*, ser. GLS'11. Pittsburgh, PA, USA: ETC Press, 2011, p. 55–63.

[20] T. Banerjee, G. Mukherjee, S. Dutta, and P. Ghosh, "A large-scale constrained joint modeling approach for predicting user activity, engagement, and churn with application to freemium mobile games," *Journal of the American Statistical Association*, 2020.

[21] T. Banerjee, P. Liu, G. Mukherjee, S. Dutta, and H. Che, "Joint modeling of playing time and purchase propensity in massively multiplayer online role-playing games using crossed random effects," *The Annals of Applied Statistics*, vol. 17, no. 3, pp. 2533–2554, 2023.

[22] B. Karmakar, P. Liu, G. Mukherjee, H. Che, and S. Dutta, "Improved retention analysis in freemium role-playing games by jointly modelling players' motivation, progression and churn," *Journal of the Royal Statistical Society Series A: Statistics in Society*, vol. 185, no. 1, pp. 102–133, 2022.

[23] Z. Yu, M. Gao, and L. Wang, "The effect of educational games on learning outcomes, student motivation, engagement and satisfaction," *Journal of Educational Computing Research*, vol. 59, no. 3, pp. 522–546, 2021.

[24] X. Chen, L. Niu, A. Veeraraghavan, and A. Sabharwal, "Faceengage: Robust estimation of gameplay engagement from user-contributed (youtube) videos," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 651–665, 2019.

[25] A. Miguel-Cruz, A. M. R. Rincon, C. Daum, D. A. Q. Torres, R. De Jesus, L. Liu, and E. Stroulia, "Predicting engagement in older adults with and without dementia while playing mobile games," *IEEE Instrumentation & Measurement Magazine*, vol. 24, no. 6, pp. 29–36, 2021.

[26] A. Z. Abbasi, M. Asif, L. D. Hollebeek, J. U. Islam, D. H. Ting, and U. Rehman, "The effects of consumer esports videogame engagement on consumption behaviors," *Journal of Product & Brand Management*, vol. 30, no. 8, pp. 1194–1211, 2021.

[27] A. Rashed, S. Shirmohammadi, I. Amer, and M. Hefeeda, "A review of player engagement estimation in video games: Challenges and opportunities," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2025.

[28] W. A. IJsselsteijn, Y. A. W. de Kort, and K. Poels, "The game experience questionnaire," Technische Universiteit Eindhoven, Eindhoven, Tech. Rep., 2013.

[29] C. Jennett, A. L. Cox, P. Cairns, S. Dhoparee, A. Epps, T. Tijs, and A. Walton, "Measuring and defining the experience of immersion in games," *International journal of human-computer studies*, vol. 66, no. 9, pp. 641–661, 2008.

[30] S. Rigby and R. Ryan, "The player experience of need satisfaction (pens) model," Immersyve Inc, Tech. Rep., 2007.

[31] H. L. O'Brien and E. G. Toms, "What is user engagement? a conceptual framework for defining user engagement with technology," *Journal of the American society for Information Science and Technology*, vol. 59, no. 6, pp. 938–955, 2008.

[32] G. Ruqeyya, T. Hafeez, S. M. U. Saeed, and A. Ishwal, "Eeg-based engagement index for video game players," in *2022 International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering (ETECTE)*, 2022, pp. 1–6.

[33] O. Pino and G. Romano, "Engagement and arousal effects in predicting the increase of cognitive functioning following a neuromodulation program," *Acta Bio Medica: Atenei Parmensis*, vol. 93, no. 3, 2022.

[34] A. Apicella, P. Arpaia, M. Frosolone, G. Improta, N. Moccaldi, and A. Pollastro, "Eeg-based measurement system for monitoring student engagement in learning 4.0," *Scientific Reports*, vol. 12, no. 1, p. 5857, 2022.

[35] A. Winklbauer, B. Stiglbauer, M. Lankes, and M. Sporn, "Telling eyes: Linking eye-tracking indicators to affective variables," in *Proceedings of the 18th International Conference on the Foundations of Digital Games*, ser. FDG '23. New York, NY, USA: Association for Computing Machinery, 2023.

[36] E. Wood, T. Baltruaitis, X. Zhang, Y. Sugano, P. Robinson, and A. Bulling, "Rendering of eyes for eye-shape registration and gaze estimation," in *2015 IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile: IEEE, 2015, pp. 3756–3764.

[37] D. Gábana Arellano, L. Tokarchuk, and H. Gunes, "Measuring affective, physiological and behavioural differences in solo, competitive and collaborative games," in *Intelligent Technologies for Interactive Entertainment*. Cham: Springer, 2017, pp. 184–193.

[38] C.-H. Wu, S.-Y. Liu, X. Huang, X. Wang, R. Zhang, L. Minciullo, W. K. Yiu, K. Kwan, and K.-T. Cheng, "Cmose: Comprehensive multimodality online student engagement dataset with high-quality labels," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024, pp. 4636–4645.

[39] S. Huynh, S. Kim, J. Ko, R. K. Balan, and Y. Lee, "Engagemon: Multimodal engagement sensing for mobile games," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 2, no. 1, pp. 1–27, 2018.

[40] K. Pinitas, D. Renaudie, M. Thomsen, M. Barthet, K. Makantasis, A. Liapis, and G. N. Yannakakis, "Predicting player engagement in tom clancy's the division 2: A multimodal approach via pixels and gamepad actions," in *Proceedings of the 25th International Conference on Multimodal Interaction*, ser. ICMI '23. NY, USA: ACM, 2023, p. 488–497.

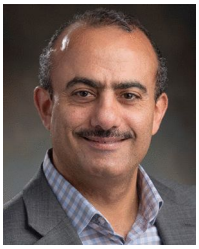
[41] D. Reguera, P. Colomer-de Simón, I. Encinas, M. Sort, J. Wedekind, and M. Boguñá, "Quantifying human engagement into playful activities," *Scientific Reports*, vol. 10, no. 1, p. 4145, 2020.

[42] A. Rashed, S. Shirmohammadi, and M. Hefeeda, "Real-time player engagement measurement using non-intrusive game telemetry," *IEEE Open Journal of Instrumentation and Measurement*, pp. 1–1, 2025.
[43] P. Buono, B. De Carolis, F. D'Errico, N. Macchiarulo, and G. Palestra, "Assessing student engagement from facial behavior in on-line learning," *Multimedia Tools and Applications*, vol. 82, no. 9, pp. 12 859–12 877, 2023.
[44] C. T. Tan, S. Bakkes, and Y. Pisan, "Inferring player experiences using facial expressions analysis," in *Proceedings of the 2014 Conference on Interactive Entertainment*, 2014, pp. 1–8.
[45] A. Denisova, A. I. Nordin, and P. Cairns, "The convergence of player experience questionnaires," in *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*. NY, USA: ACM, 2016, p. 33–37.
[46] 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," *Educational technology research and development*, vol. 72, no. 5, pp. 2585–2615, 2024.
[47] T. Baltrušaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 59–66.
[48] A. Toisoul, J. Kossaifi, A. Bulat, G. Tzimiropoulos, and M. Pantic, "Estimation of continuous valence and arousal levels from faces in naturalistic conditions," *Nature Machine Intelligence*, vol. 3, no. 1, pp. 42–50, 2021.
[49] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.
[50] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.
[51] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (eeg) classification tasks: a review," *Journal of neural engineering*, vol. 16, no. 3, p. 031001, 2019.
[52] H.-C. Chiang, Y.-H. Wu, G.-H. Li, S. Shirmohammadi, and C.-H. Hsu, "Palm: Personalized active learning for mmwave-based activity recognition," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–14, 2025.
[53] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17. JMLR.org, 2017, p. 3319–3328.
[54] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, "Captum: A unified and generic model interpretability library for pytorch," 2020.
[55] J. A. Miller, U. Narayan, M. Hantsbarger, S. Cooper, and M. S. El-Nasr, "Expertise and engagement: re-designing citizen science games with players' minds in mind," in *Proceedings of the 14th International Conference on the Foundations of Digital Games*, ser. FDG '19. NY, USA: ACM, 2019.

Ammar Rashed received his B.S. degree in Computer Science and Engineering with the highest standing in his graduating class in 2018 from İstanbul Şehir University, Türkiye, and his M.S. degree in Computer Science in 2020 from Özyeğin University, Türkiye. He completed his Ph.D. in Computer Science at the University of Ottawa, Canada, in 2025, where his work focused on real-time player engagement measurement in video games. His research contributions span machine learning, affective computing, and computational social science, with emphasis on AI-assisted measurements and behavior modeling. His doctoral research has been published in *ACM TOMM*, *IEEE OJIM*, and *IEEE Data Descriptions*, while earlier work includes unsupervised stance detection published in the *AAAI ICWSM* and collaborative research on Arabic NLP ranked first the OSACT shared task at the *ACL Workshop on Arabic NLP*. With 10+ publications, his work has garnered over 440 citations.

Shervin Shirmohammadi (M'04, SM'04, F'17) received his Ph.D. in Electrical Engineering in 2000 from the University of Ottawa, Canada, and after spending 3 years in the industry as a senior architect and project manager, joined as Assistant Professor the same University, where since 2012 he has been a Full Professor with the School of Electrical Engineering and Computer Science. He is Director of the Discover Laboratory, doing research at the intersection of AI-assisted measurements and multimedia systems and networks. The results of his research, funded by more than $28 million from public and private sectors, have led to over 450 publications, over 80 researchers trained at the postdoctoral, PhD, and Master's levels, 30+ patents and technology transfers to the private sector, and four Best Paper awards. He was an Associate Editor of the IEEE Transactions on Circuits and Systems for Video Technology from 2016 to 2019, receiving the Best Associate Editor Award in 2016 and 2018, an Associate Editor of ACM Transactions on Multimedia Computing Communications and Applications from 2009 to 2023, receiving the Associate Editor of the Year Award in 2009, 2010, 2018, and 2019, and Associate Editor of Springer's Multimedia Tools and Applications from 2004 to 2012. He has chaired and/or served on the organizing committees of many editions of premier multimedia conferences including ACM Multimedia, IEEE ICME, ACM MMSys, IEEE MIPR, and ACM NOSSDAV.

Dr. Shirmohammadi is an IEEE Fellow "*for contributions to multimedia systems and network measurements*", and recipient of the 2019 George S. Glinski Award for Excellence in Research, the 2021 IEEE IMS Distinguished Service Award, and the 2023 IEEE IMS Technical Award "*for contributions to the advancement of machine learning-assisted measurements*".



Mohamed Hefeeda (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees from Mansoura University,Mansoura, Egypt, in 1994 and 1997, respectively, and the Ph.D. degree from Purdue University, West Lafayette, IN, USA, in 2004. He is a Professor in the School of Computing Science at Simon Fraser University (SFU), Burnaby, BC, Canada, where he served as the Director of the School between 2018 and 2023. He founded and leads the Network and Multimedia Systems Laboratory (http://nmsl.cs.sfu.ca) at SFU. His research interests include multimedia systems, mobile and wireless video streaming, immersive video processing and delivery, and network systems and protocols. He has authored or co-authored more than 150 papers and multiple granted patents. Dr. Hefeeda was the recipient of the prestigious NSERC Discovery Accelerator Supplements (DAS) awards in 2011, which is granted to a select group of distinguished researchers from all Science and Engineering disciplines in Canada. His research on mobile multimedia systems has resulted in multiple patents and conference awards (e.g., ACM MMSys Best Paper, ACM Multimedia Best Demo, and IEEE Innovation Best Paper), and has been featured in several news venues, including ACM Tech News, World Journal News, and CTV British Columbia. He has served on the editorial boards of premier journals, such as the ACM Transactions on Multimedia Computing, Communications and Applications (TOMM), where he was named the Best Associate Editor in 2014, and on the organization committees and/or Co/Chaired several international conferences, such as ACM MMSys, ACM MM, IEEE ICME, and ACM NOSSDAV.

Ihab Amer is a Professor of Practice at the Department of Computer Science and Engineering, American University of Sharjah. Before that, he was a Fellow at AMD where he was responsible for directing innovative future-looking solutions. He joined AMD in 2012 as a Principle Member of Technical Staff and was responsible for designing and defining strategies of Video CODEC solutions. Prior to AMD, he worked as an Assistant Professor at the German University in Cairo for 5 years, and as a Visiting Researcher at the Multimedia Architectures Research Group of EPFL for 2 years. He has been a member of ISO/IEC MPEG, and participated in the organization of several academic and professional events such as the 20th meeting of ISO/IEC JTC 1. He served as the local co-chair of IEEE IWSOC'06 and IEEE ICMENS'06, and organized/chaired the IEEE ICIP'09 Special Session on Reconfigurable Video Coding (RVC). He has delivered many keynotes and tutorials at venues such as ACM MMSys 2023, IEEE ISCAS 2016, and IEEE ICECS 2015. Dr. Amer received his PhD from the University of Calgary, his MSc from the American University in Cairo, and his BSc from Ain Shams University. He holds numerous granted and pending patents and several prestigious awards, including the 2015 AMD Innovation Award and the ISO/IEC award for the "special contribution" as a project editor in the international MPEG-4 Part 9 standard. He has over 40 contributions to MPEG standards, and has authored/coauthored over 40 peer-reviewed conference and journal papers, with five receiving or being nominated for best paper awards.

Dr. Amer is a Senior Member of IEEE and a licensed Professional Engineer in Ontario. His research interests include machine learning (ML) for multimedia, efficient ML architectures, application-specific video coding, and low-power multimedia architectures.