Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Establishing Objective Ground-Truth for PediatricADHD Engagement: A Methodological Frameworkand Benchmark Dataset

Somayeh Malekshahi
University of Tehran
Reza Rostami
University of Tehran
Hamid Soltanian-Zadeh
hszadeh@ut.ac.ir

University of Tehran

Article

Keywords:

Posted Date: March 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8497702/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Establishing Objective Ground-Truth for Pediatric ADHD Engagement: A Methodological Framework and Benchmark Dataset

Somayeh Malekshahi[1], **Reza Rostami**[2], **and Hamid Soltanian-Zadeh**[1,3,*]

[1]School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran
[2]Department of Psychology, University of Tehran, Tehran, Iran
[3]School of Cognitive Sciences, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran
[*]hszadeh@ut.ac.ir

ABSTRACT

Detecting engagement in pediatric ADHD typically relies on subjective, biased human labeling. To address this, we introduce an objective annotation framework utilizing behavioral performance from the Integrated Visual and Auditory Continuous Performance Test (IVA-2). From 21 children, we derived a continuous Engagement Score (E_s) and windowed categorical labels, validating them against official IVA-2 clinical scales. The proposed score demonstrated high internal reliability (r = 0.84), substantial convergent validity (AUC = 0.97), and clinical sensitivity, significantly differentiating ADHD severity groups (Kruskal–Wallis, $H = 9.71$, $p = 0.046$, $N = 21$). Temporal analysis revealed a strong concurrent correlation (r = 0.57) with real-time dynamics. Benchmarking established that static Random Forest models currently outperform temporal baselines, with eye-gaze dynamics serving as the strongest unimodal predictor. We release the synchronized video features (OpenFace 2.0) and validated labels, providing a robust, privacy-preserving benchmark for objective engagement detection in clinical and educational settings.

Introduction

Engagement is a multidimensional construct encompassing behavioral, emotional, and cognitive dimensions that collectively support successful learning and sustained task involvement. Seminal work in educational psychology conceptualizes engagement as malleable and sensitive to contextual factors, making it a central target for interventions aimed at improving student motivation and academic outcomes[1]. Richer characterizations of how students behave, feel, and think during learning tasks are needed to guide the design of effective and responsive educational environments.

In school contexts, engagement functions as a protective factor, with strong associations to academic achievement, school retention, and socio-emotional well-being[2]. These relationships are especially pronounced during key educational transitions, where engagement trajectories can serve as predictors of later academic success. Importantly, engagement does not emerge in isolation: emotional and behavioral engagement are deeply embedded within contextual dynamics, including instructional quality and student–teacher relationships[3]. Understanding and measuring engagement accurately is therefore essential for evaluating interventions and for advancing research on learning and attention.

Engagement in this study is conceptualized as a multidimensional construct comprising cognitive, behavioral, and emotional dimensions. We map our objective IVA-2 performance metrics and synchronized video features to these dimensions: cognitive engagement is represented by attention-driven accuracy and eye-gaze stability; behavioral engagement is captured through inhibitory response control and head-pose orientation; and emotional engagement is inferred from response-time consistency and facial action unit dynamics. This tripartite framework enables a more granular analysis of how ADHD-related challenges manifest during a continuous performance task.

Attention-Deficit/Hyperactivity Disorder (ADHD) is characterized by pervasive challenges in sustaining attention and regulating responses, which directly impair a child's ability to maintain consistent engagement in academic and structured tasks. Despite often having normative or even above-average cognitive potential, many children with ADHD show marked discrepancies between capability and performance, largely attributable to difficulties in attentional control and inhibitory regulation[4]. Consequently, engagement is a critical therapeutic and educational target, motivating the development of adaptive technologies, from digital learning tools to therapeutic feedback systems.

Accurate engagement measurement, however, remains a significant barrier. Traditional methods rely heavily on subjective instruments such as teacher/parent rating scales or manual video coding. These methods are labor-intensive and prone to inter-rater variability, observer fatigue, and subjective bias, making it notoriously difficult to achieve reliable ground-truth

Table 1. Comparison of publicly available engagement-related datasets.

Dataset	Year	Setting	Participants	Duration	Total Minutes	Validation
DAiSEE[13,14]	2016	In-the-Wild (Lectures)	112	10 s	1500	10 Annotators/Crowdsourced
HBCU[12]	2014	In-the-Lab (Training)	34	10 s	1530	Expert Labelers
EngageWild[15]	2018	In-the-Wild (Lectures)	78	5 min	990	5 Annotators
EngageNet[8]	2023	In-the-Wild (Web-based)	127	10 s	1860	5 Annotators
Our Dataset	**Current**	**Clinical (ADHD)**	**21**	**15 s**	**262.5**	**Automatic/Objective Clinical Metrics**

labels[5]. Although multimodal affective computing research has begun to explore automated engagement detection, the reliability of such systems is constrained by the quality of the ground truth labels. Many existing datasets use annotations from non-experts (e.g., crowdworkers) who infer on-task or off-task states based solely on visible behavior—labels that do not necessarily correspond to true cognitive engagement and lack clinical grounding.

Meanwhile, broader ADHD research highlights substantial challenges in objective measurement and prediction. Students with ADHD consistently show lower school engagement and face elevated behavioral and disciplinary risks compared to peers[2]. Developing objective diagnostic tools remains difficult due to the absence of validated biological markers or pathophysiological models[6]. Nevertheless, recent datasets have attempted to bridge this gap. For example, the HYPERAKTIV dataset provides activity and heart rate measurements from ADHD and control participants, offering a platform for physiological-based prediction models[7]. Similarly, EngageNet includes 31 hours of multimodal engagement data featuring behavioral and affective cues such as action units, eye gaze, and head pose[8]. Despite their contributions, these datasets do not include clinically validated ground-truth engagement, leaving a critical gap for research requiring objective, high-fidelity labels.

A growing body of research has attempted to overcome the limitations of subjective engagement labeling by introducing objective, sensor-derived annotations. These methods infer engagement or attentional state from physiological or behavioral signals—such as eye-tracking (gaze fixation and stability), pupillometry, facial action units, head pose, or even keystroke/mouse dynamics—on the assumption that such signals reflect underlying cognitive effort[9,10]. For example, gaze-based metrics allow moment-to-moment estimation of visual attention during digital learning tasks, providing higher temporal resolution and greater consistency than human-coded labels[10]. Other multimodal approaches combine features from facial expression, head movements, and behavioral performance to automatically generate engagement annotations without relying on manual raters[11]. Despite these advances, most of these objective pipelines lack clinical validation: they are rarely anchored to a gold-standard assessment of attention or response control. This limitation means that while such methods reduce subjectivity, they may not reflect true cognitive engagement. Hence, there is a pressing need for datasets in which objective engagement labels are explicitly tied to clinically validated measures, such as a continuous performance test, to ensure that annotations represent actual cognitive engagement, not just behavioral correlates.

The development of automated engagement detection systems has historically been supported by several publicly available datasets, many of which were designed to capture affective or cognitive states under controlled conditions. Early contributions, such as the HBCU dataset[12], collected videos of 34 undergraduate students performing cognitive training tasks, with engagement labels assigned by expert observers over 10-second segments. Subsequent efforts expanded toward in-the-wild environments to incorporate real-world variability. The DAiSEE dataset[13,14] established a benchmark for the recognition of affective states (engaged, bored, confused, frustrated) from students watching lectures in unconstrained settings. Similarly, the EngageWild dataset[15] contains 5-minute videos from 78 participants recorded across diverse natural environments, annotated for four levels of engagement intensity. These datasets rely heavily on subjective annotation procedures, typically involving multiple human annotators and inter-rater agreement checks. More recently, the EngageNet dataset[8] scaled data collection to 127 participants using a web-based interactive platform, incorporating both behavioral and cognitive aspects of engagement.

While these datasets provide critical resources for general engagement, they generally focus on adult, non-clinical populations and rely on subjective, human-annotated labels. A crucial gap remains: a dataset specifically focused on pediatric clinical cohorts where engagement is objectively quantified using standardized psychological testing. Our work addresses this by grounding its engagement labels in the IVA-2 CPT. A comparative summary of existing datasets is provided in Table 1.

The goal of this study is to introduce and rigorously validate an objective methodology for annotating engagement using the Integrated Visual and Auditory Continuous Performance Test (IVA-2) as the behavioral ground truth. Unlike conventional subjective labeling, engagement here is derived directly from quantitative task performance metrics. We compute dual-level engagement indicators: a global continuous score and a local, windowed categorical label. These metrics are quantitatively validated against official IVA-2 clinical scales to ensure reliability and clinical relevance. Furthermore, each participant's IVA-2 session was paired with a synchronized video recording; we provide extracted OpenFace 2.0[16] features to enable future cross-modal analysis. The resulting dataset provides an objective, reproducible, and clinically grounded foundation for advancing research in multimodal engagement detection. Finally, we establish initial performance benchmarks using classical

and deep learning models to demonstrate the dataset's utility and to quantify the predictive power of different behavioral modalities.

The remainder of this paper is structured as follows: Section 2 (Materials and Methods) describes the computation and validation of engagement metrics; Section 3 (Results) presents the experimental findings; and Section 4 (Discussion) discusses the findings, limitations, and future research directions.

Methods

This study introduces an objective framework for annotating engagement using the Integrated Visual and Auditory Continuous Performance Test (IVA-2) as the behavioral ground truth. We describe the video collection procedure, the automatic annotation method, the structure of the resulting dataset, and the video feature extraction process.

Participants and Ethics

To evaluate the game's ability to assess attentional engagement, a field deployment was conducted in boys-only elementary schools, consistent with the national gender-segregated education policy. Participants were male students aged 6–12 years, initially pre-identified as inattentive by their teachers based on classroom behavior and academic performance. Clinical confirmation was obtained through the Integrated Visual and Auditory Continuous Performance Test (IVA-2), administered by a certified school counselor. Based on the standardized scoring protocol of the IVA-2, 4 students were classified with mild ADHD, 2 with moderate ADHD, 9 with severe ADHD, 3 as invalid due to irregular response patterns or insufficient task engagement, and 3 as showing no indications of ADHD.

Analytical Grouping: For subsequent data analysis, participants were categorized into two groups based on their IVA-2 assessment to facilitate exploratory comparisons. This grouping structure is consistent with the preliminary analysis in the source data:

- Group 0 (No/Mild ADHD): Included students with no indications of ADHD (n=3)and those with mild ADHD (n=4), for a total of $n = 7$ participants.
- Group 1 (Moderate/Severe/Invalid ADHD): Included students with moderate (n=2), severe (n=9), and invalid (n=3) IVA-2 results (n=14 total). The Invalid results were included in the primary analysis because the underlying cause (e.g., irregular response patterns) is behaviorally consistent with significant attentional challenges.

This study was conducted in accordance with the Declaration of Helsinki. All procedures involving data collection adhered to institutional ethical standards and data protection regulations. Data were collected between 13/04/2025 and 22/04/2025. Explicit informed consent was obtained from the parents or legal guardians of all participants prior to their involvement in the study. The study protocol was approved by the Research Ethics Committee of the Faculty of Psychology and Education at the University of Tehran (Approval Code: IR.UT.PSYEDU.REC.1403.127), ensuring compliance with ethical guidelines for research involving minors and vulnerable populations.

Study Protocol

Each participant completed the IVA-2 test individually during regular school hours in a supervised and controlled environment to minimize external distractions. After receiving standardized instructions administered by an expert (PhD in psychology), students performed the test on a laptop computer using a mouse for responses. The IVA-2 is a computerized continuous performance test designed to assess attention and response control through visual and auditory stimuli. The session lasted approximately 20–25 minutes, including standardized warm-up and cool-down phases.

Each session was recorded using a fixed-position camera capturing facial expressions, head pose, and gaze direction to enable subsequent multimodal analysis. All behavioral log data and videos were anonymized prior to processing to ensure participant privacy.

Dataset Statistics

The final dataset is comprised of video data from 21 subjects who completed the full IVA-2 main test.

- Total Duration: The total analyzed video footage amounts to 4.38 hours ($\approx$ 262.5 minutes) of behavioral data.
- Window Count: This duration translates into 1050 unique 15-second windows (10-trial segments) derived from the 500 trials per subject.
- Video Resolution: All videos were recorded at a resolution of 640×480 pixels and processed at a constant frame rate of 30 frames per second.

- Data Split (Cross-Validation): Due to the small cohort size, we did not employ a fixed Train/Validation/Test split. Instead, all modeling was performed using Leave-One-Subject-Out (LOSO) cross-validation to ensure that the model is never tested on data from a subject seen during training, providing the most realistic measure of generalization across individual differences.

Automatic Engagement Annotation: Framework

The IVA-2 generates several clinical reports, including the Clinical, Standard, and Detailed reports, each offering increasing diagnostic granularity. Additional reports such as the Comprehensive, Comparative, High/Low Demand, and Sensory Modality reports provide focused analyses of attention and response control. These outputs collectively offer essential information for clinical interpretation by identifying individual strengths and weaknesses. Among them, the Comprehensive Report provides the most detailed analysis, integrating all scales from the Detailed Report, additional component scale data, and interpretive text[17].

The TrialResponseData report contains the raw trial-by-trial performance data for all 500 test trials. This report provides fine-grained information beyond summary scores, including:

- Response details: Exact responses to each presented stimulus.
- Error classification: Detailed categorization of each error type (e.g., omission, commission, spurious).
- Raw data coverage: Inclusion of both Warm-Up and Cool-Down periods, providing a complete representation of the session.

This raw data enables clinicians and researchers to perform detailed, trial-level analyses of response consistency and atypical behavior patterns that are not visible in aggregated metrics. It effectively serves as a forensic diagnostic resource for evaluating the reliability and validity of IVA-2 results.

Overview of the Automatic Annotation Pipeline

We implemented a three-step pipeline for objective engagement annotation:

1. Data extraction.
2. Primitive state generation.
3. Engagement score computation and label generation.

Step 1: Data Extraction. For each participant, the TrialResponseData file was parsed and all numerical trial-level information was extracted. The data were sorted by trial number, and only main task trials (Freq and Rare) were retained, excluding Warm-Up and Cool-Down segments. Each trial record included: `Trial #`, `Type`, `Mode` (visual/auditory), `Stimulus` (target code 1 or 2), `# Clicks` (participant response: 0, 1, or 2), `R.T.` (response time), and `Error` type (Omission, Commission, Spurious, etc.).

Step 2: Primitive State Generation. Each trial was assigned an initial correctness label based on the `Stimulus`, `# Clicks`, and `Error` fields, following IVA-2 response time standards (minimum valid R.T. = 125 ms). The decision rules were:

1. If `Error` is Spurious or Zero rx, the trial is ignored.
2. If `Stimulus`=1, `# Clicks`=1, and $R.T. > 125ms \rightarrow$ Correct.
3. If `Stimulus`=2 and `# Clicks`=0 $\rightarrow$ Correct (the participants correctly realized there was no need to click).
4. If `Stimulus`=1 and `# Clicks`=0 $\rightarrow$ Omission (Incorrect).
5. If `Stimulus`=2 and `# Clicks`=1 $\rightarrow$ Commission (Incorrect).
6. All other cases $\rightarrow$ Unknown.

Trials labeled as Correct were marked as *Likely Engaged* (LE), while Omission or Commission trials were marked as *Likely Disengaged* (LD). Trials resulting in an Unknown label, or those ignored (Rule 1), were excluded from subsequent computations.

Step 3: Engagement Labeling and Score Computation. Final engagement labels were derived at two levels: (1) a continuous global score and (2) a categorical window-based label for local segments. The quantitative foundation for both is the *Engagement Score* (E_s), defined in equation (1).

Engagement Score (E_s): Concept and Computation

The Engagement Score (E_s) is a reproducible quantitative measure designed to capture attention and response control—two core constructs of engagement—based solely on IVA-2 performance data. It integrates accuracy, error frequency, and response-time stability into a single normalized metric:

$$
\begin{aligned}
A &= \text{Number of Correct trials} \\
E &= \text{Number of Commission errors} \\
O &= \text{Number of Omission errors} \\
Total &= A + E + O \\
RTVar &= \frac{\text{std(R.T.)}}{\text{mean(R.T.)}} \\
ErrorPenalty &= \frac{(w_{omission} \cdot O + w_{commission} \cdot E)}{Total} \\
RawScore &= \frac{A}{Total} - ErrorPenalty - RTVar \\
E_s &= \text{MinMaxNorm(RawScore)}
\end{aligned}
\tag{1}
$$

Each component represents a distinct dimension of engagement:

- A – Accuracy, reflecting Cognitive Engagement (correct attention-driven responses, sustained attention).
- E and O – Commission and Omission errors, reflecting Behavioral Engagement (impulsivity and inattentiveness).
- $RTVar$ – Response time variability, a critical measure of sustained attention stability, a proxy for Affective Engagement (regulation stability).
- E_s – Normalized engagement score ranging from 0 (disengaged) to 1 (highly engaged).

The score is computed in three sequential steps:

1. Penalty Term Computation: The calculation of $ErrorPenalty$ incorporates specific weights ($w_{omission}, w_{commission}$) for error types. In this study, we assign equal weight (1.0) to both Omission and Commission errors, reflecting a clinical judgment that inattention and impulsivity are equally detrimental to sustained engagement in this context. However, these weights are exposed as hyperparameters, allowing future research to adjust for different clinical criteria.

2. Raw Score Aggregation ($RawScore$): The Raw Score is the final, unnormalized metric. It is calculated by taking the proportional measure of successful responses ($\frac{A}{Total}$) and subtracting both penalty terms computed in the previous step ($RawScore = \frac{A}{Total} - ErrorPenalty - RTVar$). This aggregation ensures that the final score reflects a balance between successful execution (accuracy) and stable performance (low errors and low variability).

3. Normalization: The Raw Score is then scaled using Min-Max Normalization across the entire subject population. This standardization ensures that the final E_s is bounded between 0 and 1, making the metric comparable across different subjects and facilitating its use as a continuous ground-truth label for machine learning models.

Global Engagement Score (GE_s) and Statistical Validation

The Global Engagement Score (GE_s) is the E_s computed across all valid Freq and Rare trials in a participant's IVA-2 session. The validation plan for GE_s was designed to assess its reliability and convergent validity against 44 extracted scales from the Comprehensive IVA-2 report. These reports provide aggregated clinical summaries (e.g., global scales, quintiles, response-time distributions) that serve as coarse, orthogonal validators—not as a perfect trial-by-trial ground truth but as clinical anchors to test construct/criterion validity. We defined mappings between our computed E_s and the report's clinical constructs, which were then evaluated using statistical validation pipelines.

The validation protocol employed correlation analyses (Pearson's r and Spearman's ρ) between GE_s and all extracted official scales (e.g., Attention Quotient and Response Control Quotient).

Local Engagement Label and Score (Categorical Window)

This approach provides a high-resolution view of engagement for short, contiguous segments of the IVA-2 session, aligned with corresponding video intervals. The local label was generated using a 10-trial, non-overlapping window (e.g., trials 1-10, 11-20). This specific size ($\approx$ 15 seconds) was selected as a principled trade-off between temporal resolution and signal stability. We evaluated window sizes from 2 to 10 trials. While point-biserial correlations with global metrics were strong and consistent across most window sizes ($r \approx 0.52 - 0.58$), the 10-trial window was preferred because it provides sufficient temporal context to capture slow-building behavioral phenomena (e.g., sustained gaze aversion or fidgeting sequences) which are not observable in shorter segments. Furthermore, this duration ($\approx$ 15 s) aligns with standard window sizes in affective computing datasets (e.g., DAiSEE), facilitating the extraction of meaningful temporal dynamics for the sequential baseline models.

A categorical label (*Engaged* or *Disengaged*) was assigned to each window using a majority vote of the *PrimitiveStatus* (LE/LD) of the valid trials within it.

To validate this categorical label, we also computed a continuous Local Engagement Score for each window, using the same E_s formula (equation (1)) applied only to the trials within that window. This created a paired dataset of (categorical label, continuous score) for each $\approx$15-second segment.

The validation pipeline for these local metrics (with results presented in Section 3) was designed to assess four key properties:

- Reliability: Assessed via split-half reliability.
- Convergent Validity: Assessed by computing the Area Under the Curve (AUC) from an ROC analysis of the categorical label versus the continuous score.
- Temporal Validity: Assessed using a cross-correlation (lag) analysis to determine if the metrics were concurrent.
- Clinical Validity: Assessed by comparing metric distributions across ADHD severity groups using the Kruskal-Wallis test.

Video Feature Extraction

To preserve participant privacy, raw video data was processed locally. We utilized the OpenFace 2.0 toolkit to extract behavioral features. To ensure anonymity, biometric indicators such as raw facial texture (HOG) and absolute landmark geometry were excluded. The final public dataset consists of non-identifiable behavioral time-series data, including head pose (6 features), eye gaze vectors (8 features), and facial action unit intensities (17 features), which collectively describe cognitive, behavioral, and affective states without revealing participant identity.

The data was collected in a clinical setting with a pediatric cohort (ADHD), introducing several real-world challenges typically associated with in-the-wild datasets. Despite controlled illumination, videos contain non-frontal poses due to subject movement and fidgeting, self-occlusions (e.g., hand near face), and low signal-to-noise ratio in temporal Action Unit (AU) expressions. This inherent challenge necessitates the use of robust feature extraction methods and is a key distinction from lab-based adult datasets.

To mitigate noise introduced by non-frontal poses and partial occlusions, a preprocessing step was applied to the raw OpenFace output. Frames were filtered using the confidence column, retaining only those frames where the feature extractor reported a tracking confidence score of 0.7 or higher.

Although raw landmark coordinates were excluded from the final dataset for privacy, this filtering ensures that the derived signals including gaze vectors (cognitive components), head-pose estimates (behavioral components), and facial action unit intensities (affective-related components) are based on successful face tracking, preserving data reliability.

Baseline Models

To establish reproducible reference points for future research, we evaluated a set of baseline models spanning classical machine learning and lightweight temporal neural architectures. These baselines quantify the difficulty of frame-level facial engagement prediction and provide an initial performance benchmark for the proposed dataset.

- Classical Linear Baseline: A Logistic Regression (LR) classifier trained on the same aggregated window-level features. This serves as a baseline to determine if the relationship between behavioral features and engagement is linearly separable.
- Classical Non-Sequential Baseline: A Random Forest (RF) classifier was trained on aggregated window-level features. For each 15-second window (corresponding to 10 IVA-2 trials), we computed the mean and standard deviation of all OpenFace-derived signals, resulting in a 62-dimensional statistical representation. Random Forest is a widely used nonlinear model in behavioral signal analysis due to its robustness under small-sample and noisy conditions.

- Sequential Neural Baselines: Given that each window contains up to 450 facial frames, we evaluated two temporal architectures to capture dynamic behavioral changes:
  - LSTM: A single-layer Long Short-Term Memory network followed by a fully connected classifier. This architecture is commonly used in affective computing for modeling sequential dependencies in facial behavior.
  - TCN: A Temporal Convolutional Network using dilated causal convolutions with residual blocks, enabling modeling of long-range temporal structure without the computational overhead of recurrence.

To isolate the contribution of specific engagement indicators, we performed an ablation study in which temporal models were trained using only Pose features (Behavioral, 6 dimensions), only Gaze features (Cognitive, 8 dimensions), or only Action Units (Affective, 17 dimensions). All models were trained exclusively on the OpenFace-derived features to ensure reproducibility.

Evaluation Protocol

We used Leave-One-Subject-Out (LOSO) cross-validation to prevent subject identity leakage. Due to class imbalance, we report Macro F1-score, Balanced Accuracy, and Global ROC-AUC aggregated over all test folds. These metrics reflect both per-class performance and overall discriminative ability. Unlike some prior work, due to the high class imbalance in our clinical dataset, we prioritize Global ROC-AUC and Macro F1-score to ensure that our models are not biased toward the majority Engaged class, thus providing a more robust measure of generalized discriminative ability.

Sensitivity Analysis Framework

To address potential confounding factors related to clinically Invalid profiles, a supplementary sensitivity sub-analysis was performed. This involved repeating the validation and benchmarking pipelines on a refined sub-sample ($N = 18$) by excluding Invalid results from Group 1.

Results

The validation of our engagement metrics was performed in two stages. First, we established the descriptive statistics for our proposed metrics (e.g., E_s, error counts) based on the full dataset (N=21). Second, we assessed the convergent validity of our Global Engagement Score (GE_s) by correlating it with the gold-standard clinical scales from the IVA-2 Comprehensive Report, which was available for a subset of participants (N=18).

Descriptive Statistics of Proposed Engagement Metrics

Table 2 presents the descriptive statistics for our computed metrics, aggregated across all 21 participants. The metrics are shown for the full test (All) and for the Visual and Auditory modalities separately. Data are presented as Mean ± Standard Deviation [Range: Min–Max]. The mean global engagement score (GE_s) for the cohort was 0.568 ± 0.136 [0.253–0.771], with comparable scores across the visual (0.553 ± 0.198 [0.031–0.801]) and auditory (0.595 ± 0.137 [0.323–0.781]) modalities. The lower minimum values observed in the visual modality are consistent with the presence of severe attentional lapses in specific clinical subgroups.

Table 2. Descriptive Statistics for Computed Engagement Metrics (N=21)

Metric	All Modalities (N=21)		Visual (N=21)		Auditory (N=21)	
	Mean	(SD)	Mean	(SD)	Mean	(SD)
Corrects	393.76	(54.87)	194.76	(32.26)	199.00	(26.55)
Commissions	25.76	(18.67)	12.91	(8.87)	12.86	(10.58)
Omissions	59.29	(42.59)	31.81	(27.80)	27.48	(19.97)
RTVar	48.09	(8.50)	49.44	(17.06)	44.39	(9.31)
GE_s	**0.57**	**(0.14)**	**0.55**	**(0.20)**	**0.60**	**(0.14)**

Engagement Score (E_s) is Correlated with Gold-Standard IVA-2 Metrics

To establish convergent validity, we correlated our modality-specific GE_s scores (Visual and Auditory) with the official clinical scales from the IVA-2 Comprehensive Report (N=21). The results, summarized in Table 3, demonstrate strong, modality-specific validation.

We report Spearman's ρ for the visual modality and Pearson's r for the auditory modality, as these analyses revealed the most consistent and statistically significant relationships. The Visual GE_s showed exceptionally strong correlations with official IVA-2 visual *attention* scales, such as Visual Attention ($\rho = 0.83, p = 0.0001$) and Visual Vigilance ($\rho = 0.88, p = 1.000 \times 10^{-5}$). Notably, it did not correlate significantly with visual *response control* metrics, excepts Visual Consistency ($\rho = 0.53, p = 0.0350$), suggesting a high degree of specificity to attentional processes.

Conversely, the Auditory GE_s correlated strongly with both auditory *attention* scales (e.g., Auditory Attention, $r = 0.84, p = 3.000 \times 10^{-5}$) and *response control* scales (e.g., Auditory Response Control, $r = 0.57, p = 0.0176$; Auditory Prudence, $r = 0.63, p = 0.0090$). This indicates that our auditory engagement score successfully captures both inattention and impulsivity, which are key constructs of the IVA-2 test.

Table 3. Convergent Validity: Correlation of GE_s with IVA-2 Comprehensive Report Scales (N=21)

Modality	IVA-2 Clinical Scale	Correlation Coefficient	p-value
Part A: Visual GE_s (Visual) vs. Visual Scales (Spearman's ρ)			
Attention	Visual Attention	0.827	$0.0001***$
Attention	Visual Vigilance	0.881	$< 0.0001***$
Attention	Visual Sustained Attention	0.811	$0.0001***$
Attention	Visual Elasticity	0.808	$0.0002***$
Attention	Visual Dependability	0.554	$0.0260*$
Attention	Visual Steadiness	0.791	$< 0.0001***$
Attention	Visual Comprehension	0.849	$< 0.0001***$
Attention	Visual Acuity	0.807	$0.0002***$
Attention	Visual Stability	0.648	$0.0066**$
Response Control	Visual Response Control	0.113	0.6761
Response Control	Visual Prudence	0.229	0.3944
Response Control	Visual Consistency	0.529	0.0350*
Part B: Auditory GE_s (Auditory) vs. Auditory Scales (Pearson's r)			
Attention	Auditory Attention	0.835	$< 0.0001***$
Attention	Auditory Vigilance	0.831	$0.0001***$
Attention	Auditory Sustained Attention	0.825	$0.0001***$
Attention	Auditory Elasticity	0.903	$< 0.0001***$
Attention	Auditory Dependability	0.590	$0.0160*$
Attention	Auditory Steadiness	0.718	$0.0003***$
Attention	Auditory Comprehension	0.818	$< 0.0001***$
Attention	Auditory Focus	0.563	$0.0231*$
Response Control	Auditory Response Control	0.567	$0.0176*$
Response Control	Auditory Prudence	0.629	$0.0090**$
Response Control	Auditory Consistency	0.560	$0.0240*$
Response Control	Auditory Reliability	0.447	$0.0420*$

Significance codes: $*p < 0.05, **p < 0.01, ***p < 0.001$

Local Engagement Metrics: Reliability, Validity, and Clinical Sensitivity

Following the validation of the global score, we assessed the 10-trial windowed metrics (*Local Engagement Score* and *Local Label*) to confirm their robustness for time-series annotation. The validation, based on 815 aggregated valid windows from 21 participants, confirmed the metrics are reliable, convergent, concurrent, and clinically sensitive.

1. Reliability (Internal Consistency): The fundamental engagement metric (E_s) demonstrated excellent internal reliability. A split-half reliability test, comparing the global E_s calculated from the first half of the session against the second half, yielded a strong, highly significant correlation ($r = 0.84, p = 1.44 \times 10^{-6}$). This high stability at the global level confirms that the E_s formula captures a consistent behavioral construct, justifying its application to the shorter local windows.

2. Convergent Validity: We assessed the convergence of the continuous *Local Engagement Score* (windowed E_s) and the categorical *Local Label* (majority vote). The two metrics were found to be strongly related:

- A point-biserial correlation showed a strong, positive relationship between the score and the label ($r_{pb} = 0.57, p = 1.32 \times 10^{-71}$).
- A Receiver Operating Characteristic (ROC) analysis was performed by pooling all 815 valid windows (from all 1050) across participants. Treating the continuous *Local Engagement Score* as a classifier for the categorical *Local Label*, the analysis yielded an Area Under the Curve (AUC) of 0.97 (Figure 1). This high AUC on the aggregated dataset confirms that the local score serves as a highly accurate proxy for the categorical label at the window level.

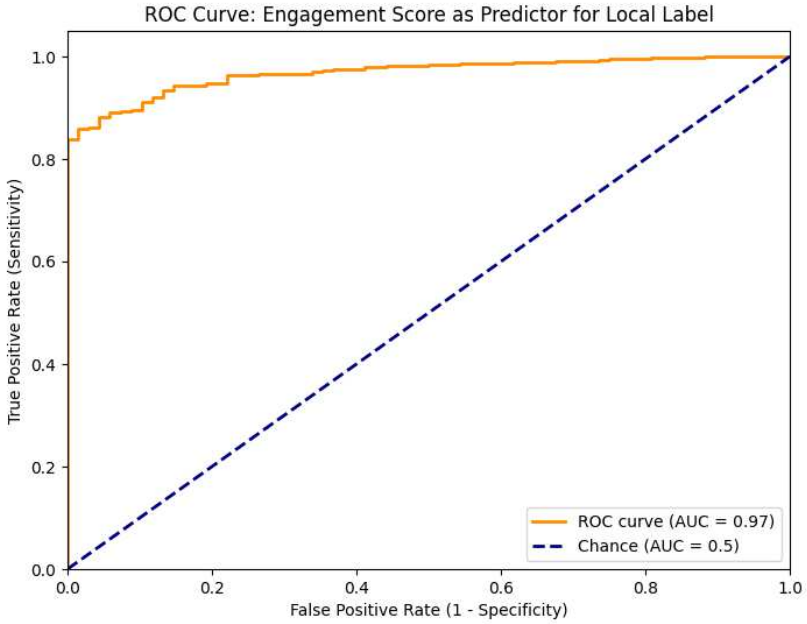


Figure 1. ROC Curve for predicting Local Label

3. Temporal Validity: A cross-correlation analysis was performed to determine if the *Local Engagement Score* was a leading or lagging indicator of the *Local Label*. The analysis revealed a peak correlation ($r = 0.57$) at a lag of 0 windows (Figure 2). This strong concurrent relationship indicates that the metrics are measuring engagement in real-time, with no significant predictive lag. This finding confirms that engagement lapses, as measured by these metrics, are rapid, in-the-moment events rather than slow-building phenomena.

4. Clinical Validity: Finally, we tested the *Local Engagement Score*'s sensitivity to clinical diagnosis. A Kruskal-Wallis test showed a statistically significant difference in mean E_s scores across the five ADHD severity groups ($H = 9.71, p = 0.046$). The group means followed the expected clinical trend: the non-ADHD group (Group 0) had the highest mean score (0.72), while the severe-ADHD group (Group 4) had the lowest (0.44).

   While the overall test was significant, post-hoc Dunn's tests with Bonferroni correction did not find significant differences between individual pairs (e.g., Group 0 vs. 4, $p = 0.085$). This is attributable to the limited sample size within the clinical subgroups ($N = 3$ and $N = 4$, respectively), which reduces the statistical power of the post-hoc comparison. However, the significant overall test and the clear, dose-dependent trend in the means support the clinical validity of the metric.

Baseline Performance

Table 4 summarizes the baseline results obtained under Leave-One-Subject-Out cross-validation. The Static Random Forest (RF) achieved the highest overall performance across all models (Macro F1 = 0.69, Global ROC-AUC = 0.62), establishing the most effective and robust benchmark. This result confirms that aggregated behavioral statistics derived from OpenFace features contain reliable discriminative information for window-level engagement prediction. The fact that RF (Non-linear) beats LR (Linear) significantly (0.62 vs 0.53 AUC) proves the relationship between features and engagement is non-linear.

In contrast, most temporal models struggled to surpass the static baseline. The LSTM trained exclusively on Gaze features achieved the highest temporal performance (Global AUC = 0.57), confirming that Gaze dynamics contain the most generalizable

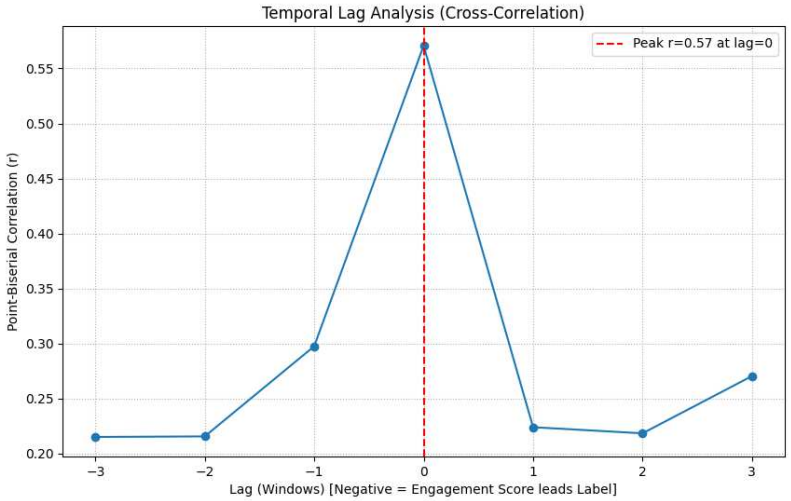


Figure 2. cross-correlation analysis to indicate the Local Engagement Score lag with Local Label

temporal signal. However, the performance of the full-feature LSTM (AUC = 0.50) and other combined-modality temporal models suggests that learning effective temporal dependencies from high-dimensional sequences remains challenging given the limited number of subjects ($N = 21$). The TCN baseline exhibited similar behavior to previous runs, achieving an F1 of 0.69 while suffering from a poor AUC (0.48), confirming its susceptibility to class imbalance. The sub-random AUC (0.44) for Pose suggests that head movement is not a reliable linear predictor of engagement in this cohort. Overall, these results provide strong evidence that simple models leveraging aggregated statistics remain the most effective baseline for small-sample behavioral engagement datasets.

Table 4. Final Baseline Model Performance (LOSO Cross-Validation) using RobustScaler and Class Weighting. The Random Forest (RF) is the strongest overall model, while Gaze is the strongest unimodal temporal predictor.

Model	Modality	Features	Bal. Acc	F1 (Macro)	Global AUC
Static Baseline					
Logistic Regression	Combined	62 (Agg.)	63	0.43	0.53
Random Forest	Combined	62 (Agg.)	92	0.69	**0.62**
Temporal Baselines					
LSTM	Combined	31	0.58	40	0.50
TCN	Combined	31	0.92	0.69	0.48
LSTM Modality Ablation					
LSTM (Pose Only)	Pose	6	0.49	0.36	0.44
LSTM (Gaze Only)	Gaze	8	0.60	0.42	**0.57**
LSTM (AUs Only)	AUs	17	0.54	0.38	0.50
LSTM (Pose+Gaze)	AUs	14	0.55	0.40	0.55
LSTM (Gaze+AUs)	AUs	25	0.58	0.41	0.51
LSTM (Pose+AUs)	AUs	23	0.54	0.38	0.52

Temporal Dynamics and Engagement Drift

The sliding-window analysis (50 windows; each window = mean of 10 consecutive trials) revealed fluctuating engagement scores across the IVA-2 session. Although short-term variability was substantial, the regression analysis revealed a negative slope coefficient, suggesting a gradual reduction in engagement as the session progressed (slope = -0.0014, p = 0.264), Figure 3.

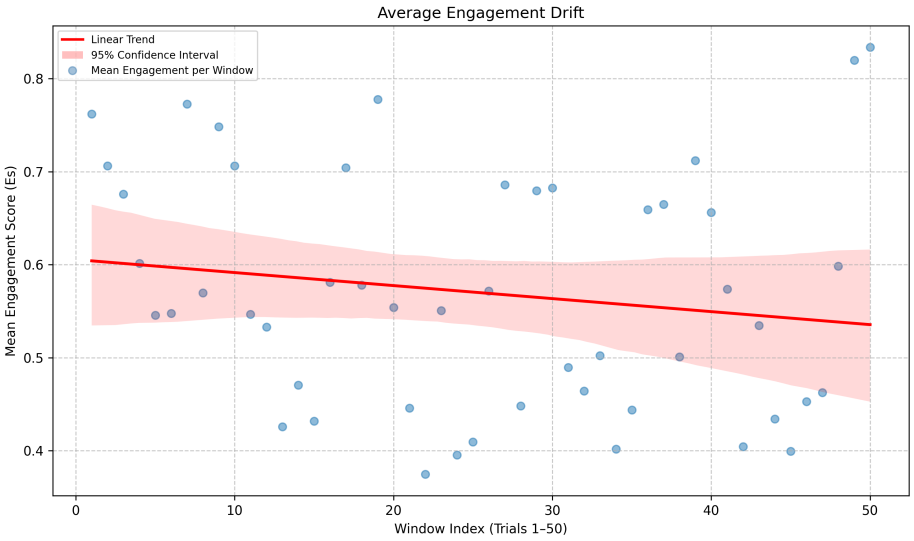


Figure 3. Average Engagement Drift across All Participants

Supplementary Analysis: Sensitivity and Robustness

To evaluate whether the inclusion of participants with Invalid IVA-2 profiles influenced the primary findings, we performed a sensitivity analysis on a refined cohort ($N = 18$; excluding Invalids from Group 1). The results, summarized in Table 5, demonstrate that the Engagement Score (E_s) remains a stable and valid metric despite the removal of non-standard response profiles.

Table 5. Descriptive Statistics for Computed Engagement Metrics (N=18), excluding Invalid participants

Metric	All Modalities (N=18)		Visual (N=18)		Auditory (N=18)	
	Mean	(SD)	Mean	(SD)	Mean	(SD)
Corrects	403.83	(51.56)	199.67	(31.55)	204.17	(24.71)
Commissions	27.67	(19.35)	13.50	(9.33)	14.17	(10.82)
Omissions	48.94	(35.31)	26.83	(26.36)	22.11	(14.76)
RTVar	47.59	(9.05)	49.38	(18.31)	43.58	(9.85)
GE_s	**0.59**	**(0.13)**	**0.57**	**(0.21)**	**0.62**	**(0.13)**

The supplementary analysis confirmed that all key psychometric properties remained robust:

- Convergent and Internal Validity: Modality-specific GE_s correlations with IVA-2 scales remained strong and consistent with the full-sample analysis (Visual Attention: $\rho = 0.83$; Auditory Attention: $r = 0.84$). The internal reliability ($r = 0.84, p = 1.17 \times 10^{-5}$) remained unchanged, proving that the metric's consistency is inherent to the E_s framework rather than the cohort composition.
- Criterion and Temporal Validity: Both the point-biserial correlation ($r_{pb} = 0.53, p = 1.56 \times 10^{-53}$) and ROC-AUC (0.97) confirmed that the discriminative power of the E_s is resilient to clinical outliers. The peak temporal correlation remained at lag 0 ($r = 0.53$), reaffirming the concurrent validity of our automated labels.
- Clinical Sensitivity: While the Kruskal–Wallis test p-value shifted slightly ($H = 6.96$, $p = 0.073$)due to reduced statistical power ($N = 18$), the clinical gradient was preserved, with the Severe group consistently exhibiting the lowest engagement scores.
- Enhanced Clinical Alignment: Notably, the exclusion of invalid profiles improved the alignment with standard scales, as the correlation between GE_s and Visual Reliability transitioned to statistical significance ($\rho = 0.49, p = 0.039$).

The temporal benchmarking results also showed improvement in the refined cohort. The Gaze-only LSTM achieved a Global AUC of 0.63, matching the performance of the combined-feature LSTM and establishing eye-tracking dynamics as the primary unimodal predictor of engagement lapses in this clinical setting (Table 6).

Table 6. Final Baseline Model Performance (LOSO Cross-Validation) using RobustScaler and Class Weighting. The LSTM is the strongest overall model, while Gaze is the strongest unimodal temporal predictor. (N=18), excluding Invalid participants

Model	Modality	Features	Bal. Acc	F1 (Macro)	Global AUC
Static Baseline					
Logistic Regression	Combined	62 (Agg.)	68	0.43	0.48
Random Forest	Combined	62 (Agg.)	94	0.73	0.58
Temporal Baselines					
LSTM	Combined	31	0.71	47	**0.63**
TCN	Combined	31	0.94	0.73	0.33
LSTM Modality Ablation					
LSTM (Pose Only)	Pose	6	0.51	0.36	0.42
LSTM (Gaze Only)	Gaze	8	0.67	0.45	**0.63**
LSTM (AUs Only)	AUs	17	0.54	0.38	0.50
LSTM (Pose+Gaze)	AUs	14	0.70	0.49	0.59
LSTM (Gaze+AUs)	AUs	25	0.75	0.49	0.62
LSTM (Pose+AUs)	AUs	23	0.69	0.44	0.51

Discussion

We developed and validated an objective framework for labeling engagement in children with ADHD using data derived from the IVA-2 continuous performance test. The proposed Engagement Score (E_s) and its corresponding windowed labels demonstrated strong reliability, clinical validity, and convergence with official IVA-2 measures.

Validating E_s against official IVA-2 clinical scales established a quantitative and clinically interpretable foundation. This validation justifies the application of our metrics to shorter video segments, bridging behavioral test data with multimodal video analysis.

Engagement, as a dynamic and multidimensional construct, cannot be adequately represented by a single label for an entire 12.5-minute session or by per-trial decisions alone. Our approach, therefore, applies a windowed labeling scheme, in which the number of trials per window (x) acts as a hyperparameter. This study empirically validated a 10-trial window, which was found to provide a robust trade-off between temporal resolution and label stability.

The obtained AUC (0.97) indicates that the categorical Local Label serves as a highly accurate proxy for the continuous E_s, confirming strong convergent validity. The zero-lag correlation between the performance metrics and the derived label suggests that attentional lapses occur almost instantaneously, reflecting concurrent cognitive, behavioral, and affective fluctuations. Additionally, the Kruskal–Wallis analysis showed significant group-level differentiation, confirming that the proposed engagement metric captures clinically meaningful differences rather than arbitrary fluctuations.

The observed pattern suggests that participants maintained moderate engagement during early segments, followed by a trajectory of reduced sustained engagement during later windows. While the slope did not reach statistical significance, the directionality is consistent with expected attentional fatigue during continuous performance tasks. The oscillatory pattern overlaid on the declining trend indicates intermittent fluctuations in momentary attentional effort, underscoring the sensitivity of the engagement score to cumulative temporal demands. This supports the use of sliding-window metrics as a fine-grained indicator of attentional stability over extended sessions.

The baseline modeling experiments provided critical insights into the limitations and strengths of the dataset. The superior performance of the Static Random Forest (AUC 0.62) over all deep temporal models highlights the challenge posed by the small subject count ($N = 21$). The complexity introduced by temporal architectures, combined with the limited training data, prevents effective generalization, reinforcing the need for feature engineering (aggregation) over complex modeling in this small-data regime.

The ablation study confirmed the clinical relevance of Gaze features, which established the highest performing temporal baseline (AUC 0.57). This result aligns with the visual nature of the Continuous Performance Test (CPT) and confirms that visual inattention is the most generalizable behavioral marker captured temporally. Conversely, the poor performance of the AUs-Only (AUC 0.50) and Pose-Only (AUC 0.44) LSTMs suggests that temporal variations in facial expressions and head movements are largely non-generalizable noise. The significant drop in performance when combining features further confirms the challenge of multimodal fusion on this dataset.

The overall results establish that the dataset presents a robust, challenging benchmark, indicating that future research must prioritize techniques such as transfer learning, subject-level pre-training, or advanced feature selection to effectively utilize the

full temporal, multimodal feature set.

A significant finding in our benchmarking was the poor generalizability of Head Pose (AUC = 0.44) compared to Eye Gaze (AUC = 0.57). This discrepancy likely reflects a subtype bias within the cohort. Our participants consisted of combined (ADHD-C) and inattentive (ADHD-I) types, excluding purely hyperactive-impulsive (ADHD-HI) individuals. Children with ADHD-I often manifest quiet disengagement, characterized by physical stillness (low pose variance) despite internal cognitive lapses and high omission rates.

Furthermore, in pediatric ADHD, fidgeting and head movement (high pose variance) are non-linear signals; they can occur both during high engagement as a self-regulation mechanism and during disengagement. In contrast, Eye Gaze remains more directly anchored to the visual attention demands of the CPT. This indicates that for inattentive pediatric cohorts, gaze dynamics are a far more reliable temporal biomarker of engagement than overt physical movement.

Existing ADHD-related video datasets primarily rely on subjective expert ratings (e.g., on-task vs. off-task), which suffer from inter-rater variability and lack a clinical reference. Our dataset's primary distinction is its objectively-derived ground truth. By anchoring engagement labels to performance on a standardized neuropsychological test (IVA-2), we bypass subjective interpretation. This design bridges the gap between clinically validated constructs (i.e., inattention, impulsivity) and observable multimodal data, offering a more robust benchmark for training and evaluating automated engagement recognition systems.

Beyond clinical alignment, the proposed annotation framework introduces several methodological advantages that distinguish it from existing subjective approaches.

- Objective Validation: Engagement indicators are directly comparable with official IVA-2 clinical scales, enabling reproducible, quantitative validation independent of human raters.

- Clinical Alignment: Labels are grounded in clinically recognized neuropsychological constructs—accuracy, error type, and response stability—ensuring interpretability within neuropsychological frameworks.

- Multimodality: The resulting dataset supports synchronized analysis of visual and behavioral, facilitating cross-domain research in attention modeling and behavioral assessment.

- Non-intrusive Methodology: Unlike engagement metrics dependent on specialized hardware (e.g., eye-trackers, EEG), restricting their use to laboratory settings, our framework validates a hardware-independent metric (E_s) derived solely from standard behavioral telemetry. This ensures the method is scalable and deployable in naturalistic settings (schools, home care) using standard webcams and input peripherals.

Despite these promising results, several limitations warrant consideration. The primary limitation of this study is the modest sample size (N = 21), which restricts the statistical power for fine-grained subgroup comparisons. While the observed trends were consistent with theoretical expectations and clinical grading, larger samples are required to establish robust generalizability.

Second, the cohort was restricted to male participants. While this decision was driven by local educational settings, it also served to maximize internal validity by controlling for the high phenotypic heterogeneity between sexes in ADHD. However, girls with ADHD often manifest the inattentive subtype, which is frequently under-identified due to a lack of overt hyperactive-impulsive behavior. Because our objective framework focuses on gaze and performance dynamics (E_s) rather than subjective behavioral impressions, it is theoretically well-suited for identifying these quiet forms of disengagement. Future studies should prioritize mixed-gender cohorts to explicitly evaluate the framework's sensitivity to female-specific phenotypic expressions.

Third, the modest AUC scores (0.50–0.63) for the temporal baselines highlight the inherent complexity of pediatric digital phenotyping. Unlike adult populations, children with ADHD exhibit masked disengagement, where they may maintain screen-directed gaze (Cognitive) while experiencing an internal attentional lapse (Affective). This baseline performance establishes a Complexity Gap, confirming that simple, off-the-shelf temporal architectures are insufficient to capture the non-linear relationship between facial micro-expressions and objective cognitive states in this population.

Consequently, future work will leverage this dataset to investigate multimodal fusion strategies that integrate facial dynamics with the validated IVA-2 ground-truth labels. Additionally, the real-time formulation of the E_s metric provides a methodological foundation for future Adaptive Digital Therapeutics. By embedding objective engagement estimates into interactive cognitive tasks, it becomes possible to move toward closed-loop systems that dynamically adjust task parameters—such as difficulty or reinforcement timing—in response to moment-to-moment engagement fluctuations. While these applications are beyond the scope of the present study, the proposed framework and dataset offer a principled, clinically grounded basis for their systematic investigation.

Conclusion

This study introduced a novel methodology for objective and clinically validated engagement annotation in children with ADHD. By leveraging IVA-2 trial-level behavioral data, the proposed Engagement Score and its categorical windowed extension enable granular, reproducible labeling of engagement states. Furthermore, we established initial performance benchmarks, demonstrating that the static Random Forest is the most robust overall baseline, while temporal gaze dynamics serve as the strongest unimodal temporal predictor of engagement in this cohort. To the best of our knowledge, the resulting dataset constitutes one of the first clinically grounded benchmark for automated engagement detection, supporting the development of reliable, interpretable machine learning models in educational and clinical settings. Future research is required to explore unsupervised pre-training and multimodal fusion techniques to further enhance the generalizability of engagement biomarkers in high-variability clinical cohorts.

Data Availability

Due to participant privacy and ethical constraints regarding facial reconstruction, the raw video recordings cannot be shared. All data generated or analyzed during this study (processed features and clinical labels) are deposited in the following repository: https://figshare.com/s/b527b0f384cce9d462a3. This dataset includes:

- Features: Non-identifiable time-series data extracted by OpenFace 2.0, comprising Head Pose (6 features), Eye Gaze vectors (8 features), and Facial Action Unit intensities (17 features).
- Ground Truth Labels:
    1. Primitive Labels: Trial-by-trial engagement status and correctness data, allowing researchers to reconstruct windows of any duration.
    2. Recommended Labels: The validated 10-trial windowed (*Final Label*) and continuous *Local Engagement Score* used in this study.
- Clinical Validation Data: Anonymized scales extracted from the IVA-2 Comprehensive Reports and raw data extracted from Trial Response Data Reports.

Code Availability

The custom Python scripts used for data alignment, window generation, and baseline modeling are available in the repository linked in the Data Availability statement.

References

1. Fredricks, J. A. *et al.* School engagement: Potential of the concept, state of the evidence. *Rev. Educ. Res.* **74**, 59–109 (2004). Accessed 15 Nov. 2025.
2. Zendarski, N., Sciberras, E., Mensah, F. & Hiscock, H. Early high school engagement in students with attention/deficit hyperactivity disorder. *Br. J. Educ. Psychol.* **87**, 127–145, DOI: 10.1111/bjep.12140 (2017).
3. Rushton, S., Giallo, R. & Efron, D. Adhd and emotional engagement with school in the primary years: Investigating the role of student-teacher relationships. *Br. J. Educ. Psychol.* **90**, 193–209, DOI: 10.1111/bjep.12316 (2020).
4. Barkley, R. A. *Attention-Deficit Hyperactivity Disorder: A Handbook for Diagnosis and Treatment* (Guilford Press, 2018), 4 edn.
5. Whitehill, J., Wu, T.-D., Bergsieker, A. & Movellan, J. R. The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affect. Comput.* **5**, 410–423 (2014).
6. Olivetti, E., Greiner, S. & Avesani, P. Adhd diagnosis from multiple data sources with batch effects. *Front. Syst. Neurosci.* **Volume 6 - 2012**, DOI: 10.3389/fnsys.2012.00070 (2012).
7. Hicks, S. A. *et al.* Hyperaktiv: An activity dataset from patients with attention-deficit/hyperactivity disorder (adhd). In *Proceedings of the 12th ACM Multimedia Systems Conference*, MMSys '21, 314–319, DOI: 10.1145/3458305.3478454 (Association for Computing Machinery, New York, NY, USA, 2021).
8. Singh, M. *et al.* Do I have your attention: A large scale engagement prediction dataset and baselines. ICMI '23, 174–182, DOI: 10.1145/3577190.3614164 (Association for Computing Machinery, New York, NY, USA, 2023).

9. D'Mello, S. & Graesser, A. Dynamics of affective states during complex learning. *Learn. Instr.* **22**, 145–157, DOI: https://doi.org/10.1016/j.learninstruc.2011.10.001 (2012).

10. Hutt, S. *et al.* "out of the fr-eye-ing pan": Towards gaze-based models of attention during learning with technology in the classroom. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, UMAP '17, 94–103, DOI: 10.1145/3079628.3079669 (Association for Computing Machinery, New York, NY, USA, 2017).

11. Yan, L., Wu, X. & Wang, Y. Student engagement assessment using multimodal deep learning. *PLOS ONE* **20**, 1–19, DOI: 10.1371/journal.pone.0325377 (2025).

12. Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A. & Movellan, J. R. The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affect. Comput.* **5**, 86–98, DOI: 10.1109/TAFFC.2014.2316163 (2014).

13. Gupta, A., D'Cunha, A., Awasthi, K. & Balasubramanian, V. Daisee: Towards user engagement recognition in the wild. *arXiv preprint* DOI: 10.48550/arXiv.1609.01885 (2016). arXiv:1609.01885.

14. Kamath, A., Biswas, A. & Balasubramanian, V. A crowdsourced approach to student engagement recognition in e-learning environments. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)* (2016).

15. Kaur, A., Mustafa, A., Mehta, L. & Dhall, A. Prediction and localization of student engagement in the wild. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*, 1–8, DOI: 10.1109/DICTA.2018.8615851 (2018).

16. Baltrusaitis, T. & et al. Openface 2.0: Facial behavior analysis toolkit. In *Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition*, 59–66, DOI: https://doi.org/10.1109/FG.2018.00019 (IEEE, Xi'an, China, 2018).

17. BrainTrain. What reports are available for iva-2? (2025).

Acknowledgments

The authors express their sincere gratitude to Dr. Mohammad Khaksar (PhD in psychology) for their invaluable contribution to this study, specifically in the recruitment of participants and the administration of the IVA-2 Continuous Performance Tests. His clinical expertise was instrumental in ensuring the validity of our cohort. Additionally, the authors gratefully acknowledge Javad M. Kheyridoost for his extensive technical guidance and multidisciplinary expertise in data science and system design. His valuable discussions informed key aspects of the data annotation framework and technical validation, contributing significantly to the overall rigor of this work.

Author Contributions

S.M. conceived the study, developed the methodology and software, performed data curation and formal analysis, and wrote the original draft. R.R. advised on data curation protocols, validated the labeling methodology against clinical reports, and reviewed the manuscript. H.S.Z provided computational resources, supervised the project, and reviewed the manuscript. All authors read and approved the final manuscript.

Funding

This study received no funding.

Competing interests

The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryDataS1Code.zip
- SupplementaryDataS2Dataset.zip