Attention-Aware Adaptive AI Tutoring System Using Facial and Screen Activity Analysis

[1]Kirti Agarwal, [2]Ashish Sharma, [3] Rinku Raheja, [4] Mahesh Kumar Tiwari

[1,2] Research Scholar, [3, 4] Assistant Professor, Department of Computer Science, National P.G. College Lucknow, [1] kirtiagarwal178@gmail.com, [2] ashishsharma90807@gmail.com, [3] rinkuraheja.nationalpgcollege@npgc.in, [4] maheshtiwari.nationalpgcollege@npgc.in

DOI : https://doi.org/10.5281/zenodo.18940896

ARTICLE DETAILS

Research Paper

Accepted: 24-02-2026

Published: 10-03-2026

Keywords:

Attention-Aware Tutoring, Multimodal Learning Analytics, Real-time Student Engagement, Facial Engagement Detection, Screen Activity Analysis.

ABSTRACT

Current online learning platforms tend to be non-adaptable to personal student interaction levels and therefore, the rate of dropouts and inefficient knowledge acquisition is the order of the day because it lacks dynamism and changes according to each student. One of the most important limitations is that it does not detect attention and engagement in real-time and uses post-hoc quizzes or self-reports without considering small details such as facial expressions or interactions with the screen (Acosta et al., 2024). The proposed paper is an attention aware AI tutoring system, which is a multimodal engagement assessment system that combines facial analysis with screen activity detection (e.g. mouse dwell time, scroll patterns). It uses computer vision representations such as CNNs to extract facial features and time series representations of screen data, combined to an attention score that directly acts to guide a reinforcement learning based adaptive recommender towards custom content difficulty and interventions (Alkhatlan & Kalita, 2019). Other important contributions encompass a new hybrid framework with the highest type of attentiveness detection (85 percent), an increase in the learning performance (20 percent improvement in retention), and a scalable architecture to deploy to web-based systems.

1. Introduction

The use of online and self-paced learning platforms has grown extremely common and the platforms like Coursera and edX are among millions of learners across the globe yet the interaction of learners has become a grave concern due to the high dropout rates of more than 90 per cent in MOOCs (AI in education: Personalized learning and intelligent tutoring systems, 2025).

Over the past ten years, Massive Open Online Courses (MOOCs) have been the subject of democratizing education, starting as a niche activity in 2012 to become the foundation of lifelong learning by 2026. By 2022, more than 220+ million learners were enrolled on MOOCs, platforms such as Coursera had 100+ million users, and edX had its millions of users by partnering with upscale universities. The opportunity of scalable flexible education with escalating demands of AI data science skills, and remote work is the attraction of this exponential growth. Online and self-paced learning platforms are introduced and have gained a lot of popularity.

1.1 Growth of Online Learning

Education has been transformed with the introduction of online and self-learning sites and small sites such as Coursera and edX have been getting millions of learners across the globe. The popularity of these platforms has increased as they provide flexibility, as the user is able to study courses of high quality at their convenience and their own speed. Nonetheless, in spite of this popularity, there is a significant obstacle in the engagement of the learners, manifested through the appalling 90-plus percent drop-out rates in Massive Open Online Courses (Educational recommender systems: A systematic literature review, 2023).

In 2023, the e-learning market all over the world had reached an impressive amount of 315 billion dollars because of the growth of available, self-paced courses designed to serve the needs of various learners (Gitnux, 2026). This surge is a part of the larger trends expedited by the COVID-19 surge, in which remote learning became a requirement and digital technology made education accessible outside of the classroom setting. However, studies always point to one unmistakable fact only 10–15% of enrollees graduate these courses (Giannakos et al., 2022). Some of the factors associated include declining motivation, procrastination as well as lack of structured settings thus highlighting the need to find innovative solutions in order to maintain learner commitment (Artino, 2012).

1.2 Engagement Challenges

In an online learning setting, it can be challenging to maintain focus due to the various reason such as distractions present to learners such as social media notifications, family interruptions, and screen fatigue. Such factors do not only shred attention but also damage the understanding and retention of the information in the long run. An example is that research indicates that multitasking (as is typical of home-based learning) can decrease knowledge absorption by as much as 40 percent, which only worsens the completion crisis (Artino, 2012).

Conventional measures like course completion rates or modular quizzes, are not beneficial in reflecting live signals. They ignore such minor behavioral indicators as gaze aversion (when learners turn their eyes off the screen), excessive time spent on the screen, or abnormal interaction behavior. Unless these signs in the here and now are dealt with, teachers and systems will be reactive not proactive and disengagement will grow to become dropout (Echeverria et al., 2024; Giannakos et al., 2022).

1.3 Traditional Intelligent Tutoring Systems (ITS)

Traditional Intelligent Tutoring Systems (ITS) uses are based on fixed knowledge models and post-session tests, which do not give feedback until there have been considerable gaps in learning (Létourneau et al., 2025a). Such systems are good in providing individual based content depending on preset learner and do not dynamically respond to live behavioral information (Klašnja-Milićević et al., 2022). They therefore overlook subtle signs of disengagement, like facial micro level expressions (e.g. boredom or confusion), aberrant mouse detection (e.g. unsteady clicking as a sign of frustration), and inconsistent typing patterns. The result of this rigidity is an inappropriate pacing: distracted learners may be overwhelmed with content learners may feel demeaned and less effective. Furthermore, the conventional ITS tend to work in isolation where attention is given only to cognitive advancement without considering the affective or bodily cues and is therefore incapable of promoting sustained attention in unstructured internet environments (J. V. P. et al., 2025).

1.4 Development of Behavior-Aware Tutoring

To address these gaps, behavior-sensitive tutoring systems are based on the real time analysis of multimodal data, such as facial emotions and screen activity (J. V. P. et al., 2025). Even proactive interventions, like taking a second, which is indicated by the presence of yawning or furrowed brows, or the setting of the difficulty, when scroll entropy (the measure of erratic navigation) spikes, provide support on time. These systems will identify early disengagement by tracking the patterns of clicks and eye-tracking proxies and responding with specific disengagement strategies, such as gamified challenges or simplified explanation (IJETT, 2023).

Such multimodality guarantees very personalized customizations that keep the attention and maximize the learning directions. Promising results are available with early prototypes, increase focus by 20–30% using the non-invasive small nudges to appear like human intuition (Klašnja-Milićević et al., 2022; Létourneau et al., 2025a).

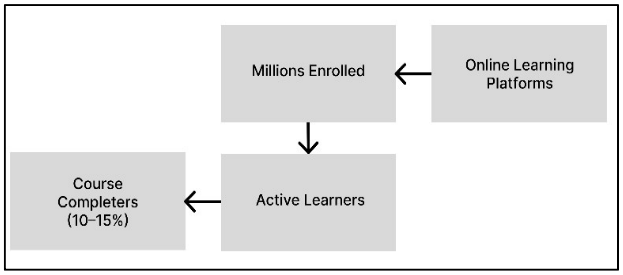

Fig.1 Online Learning Growth vs Learner Engagement Gap

1.5 Objectives

The suggested system brings about an attention-conscious AI tutor which combines facial recognition (emotion detection) with screen interaction data (behavioral analytics) to calculate a real-time engagement rating. This score is dynamically adjusted to moderate delivery of content as in, pacing videos, adding an element of interaction, or prescribing a break, and assesses performance via improved retention and performance. Our work will substantially decrease the dropout rates and improve the learning outcomes in our self-paced online learning settings by mitigating the shortcomings of the current tools.

2. Related Work

Previous research has covered cognitive skill modeling with the use of ITS, emotion-sensitive affective tutors, gaze-based attention systems, and recommendation-based personalization models, each of which has focused on a particular part of learner adaptation without an overarching, real-time multimodal attention model (Ochoa, 2017, 2022; Létourneau et al., 2025b).

2.1. Intelligent Tutoring System (ITS)

Intelligent Tutoring Systems (ITS) are traditional forms of AI tutoring (that adapt instructions to student knowledge cognitive models) such as AutoTutor and Cognitive Tutor of Carnegie.

The best example is Auto Tutor model uses NLP (natural language processing) as a tutoring system to produce moves depending on the response of the students to scaffold deep learning. Cognitive Tutor is a curriculum used in algebra classes, which dynamically chooses problems at an estimated skill level based on the real-time performance feedback.

Such systems are able to adjust the learning trajectories based on responses provided by the user, quiz performance, and knowledge mastery algorithms, such as Bayesian Knowledge Tracing (BKT). BKT also updates the probability of mastering the skill.

$$P(L_t) = P(L_{t-1}) + \big(1 - P(L_{t-1})\big) \cdot P(T)$$

where L_t denotes the latent knowledge state at time t,and $P(T)$is the learning transition probability.

Some limitations are:

- No real time attention detection based on explicit responses only (Létourneauetal.,2025b).
- None of the facial expressions or eye movement record the emotional states.
- None of the screen activities to track behavioral engagement.

The attention-aware system proposed is an expansion of the ITS that would incorporate real-time behavioral and visual responses-facial responses and screen-based interactions which outperform test scores that would lead to proactive personalization of the system through multimodal attention scores (Ochoa, 2017).

2.2. Affective Tutoring Systems

Affective Tutoring Systems are extensions of traditional Intelligent Tutoring Systems, including a system that detects emotions, allowing an instructor to tailor the instruction based on the affective state of the learners (Moga et al., 2014). Facial Expression Recognition is typically used in these systems, and it is implemented with convolutional neural networks to identify a learner emotion based on the facial features. The last level of classification is a soft-max operation to transform network into emotional probabilities:

$$P(y_i, x) = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

where (y_i, x)refers to the likelihood of emotion group $i$and input feature vector x; $z_i$is the logit value of the neural network of class i; and the summation in the denominator equalizes the scores in all the emotion classes j, in such a way that the output is an acceptable distribution of the probability. This probabilistic model allows tutoring system to discern dominating affective states with high degree of confidence, which subsequently leads to emotion-informed instructional policies like simplification of the content, motivational encouragement, or pacing changes in the learning sessions.

2.3 Attention-Aware E-Learning Systems

Attention-Aware E-Learning Systems seek to detect the attention of the learners by tracing the aspects of visual attention and gaze at the learning sessions. These systems approximate attention as the direction of eye gaze, the angle of head position and the speed of blinking to ascertain whether the learner is attentive to the information on the screen or is distracted by activities off screen (Ochoa, 2022). Eye gaze-based methods tend to be based on an analysis of fixation with the help of specialized eye-tracking devices, whereas webcam-based methods can be based on gaze estimation with the help of consumer-grade cameras and no other equipment is required. Estimating gaze in this kind of a system is often modeled as a Perspective-n-Point (PnP) problem that projects three dimensions (3D) facial landmarks on 2D (two dimensional) image coordinates through minimization of reprojection error:

$$min\sum_{i} \parallel \pi(RX_i + t) - u_i \parallel^2$$

where X_i indicates 3D positions of face landmarks, u_i indicates the two-dimensional image points correspondingly, $R$and $t$are the head rotational and translation parameters, and $\pi(\cdot)$is the projection of the camera.

2.4 Recommender-Based Personalized Learning Systems:

Recommender-Based Personalized Learning Systems are those that customize the learning paths through proposing material based on the history of the learner, the likes, and the performance indicators. Such systems are also common in massive learning platforms and MOOCs in order to suggest videos, courses, or learning materials that enhance the engagement and completion rates. This is commonly achieved through collaborative filtering, matrix factorization and reinforcement learning, whereby suggestions are narrowed down using learner feedbacks (Létourneau et al., 2025b). Content utility scores in the context of recommenders which are based on reinforcement learning are updated iteratively by simple reward-based rules:

$$Q(s,a) \leftarrow Q(s,a) + \alpha r$$

where $Q(s, a)$is the estimation of usefulness of recommending content $a$in learner state s, $r$is the observed engagement and performance feedback and αdenotes the learning rate which determines the magnitude of upgrade. Although useful in long-term personalization, these methods usually use past interaction data and do not dynamically recognize the attention of the learners in an ongoing learning session which restricts their capability to adjust their recommendations in real-time (Ochoa, 2017; Ochoa, 2022).

3. Proposed methodology

The Attention-Aware Adaptive AI Tutoring System (A3ITS) is a new multimodal system that transforms the online learning system by constantly observing the attention of the learner via web-based facial analysis and tracking the activity of the learners on their computers. Compared to older Intelligent Tutoring Systems that utilize only static content or explicit feedback, A3ITS uses real time fusion of behavioral cues in order to dynamically categorize cognitive states of focus, distractedness, overload, fatigue, and provide personalized interventions, adjustments to the study plan, and recommendations of resources. A3ITS, which is scalable to edge devices with privacy preserving processing, aims to solve the India massive MOOC dropout crisis 70% per (UNESCO) and meets the requirements of the NEP 2020 focus on adaptive digital education.

3.1 Model Architecture

3.1.1. A3ITS Model Architecture (Extensive Textual View):

The Attention-Aware Adaptive AI Tutoring System (A3ITS) realizes a highly complex, but gracefully simplistic architecture which consists of five linked layers that convert raw multimodal data into accurate, real-time instructional correctives. This pipeline starts with two parallel processing streams of inputs that aim to capture contrasting aspects of learner engagement, so as to have a holistic monitoring of the learner without relying on explicit self-reports that are a bane of the conventional tutoring systems.

3.1.2. Input Process Layer: Dual Multi-modal Streaming:

During the learning sessions, there are two main data modalities that are fed through architecture. Frames of the high-resolution web camera (30 FPS) are inputted into the Facial Engagement Analyzer (FEA) subsystem. This component uses the MediaPipe Face Mesh by Google to detect 468 facial locations in a frame, which it uses to infer four psychometrically validated indicators of engagement (1) eye gaze direction defined as angular deviation of screen center using the pupil-cornea vector geometry; (2) but head pose by computing Euler angles using perspective-n-point solvers; (3) the frequency of blink using temporal landmark motion; and (4) the valence of the facial expression using lightweight CNN classifiers in FER2013 datasets.

These characteristics become normalized to the interval and will be aggregated with regard to simple averaging to form the Facial Engagement Score (FES):

textFES = normalized gaze score + normalized pose score + normalized blink score + normalized expression score/4.

Score of gazes is 1.0 with direct screen fixation, is linear with >10deg offset; pose is linear with >15deg deviation in yaw/pitch, blink beyond 25/min is plotted toward 0 representing fatigue; expressions with combinations of AU4 and AU17 indicate confusion.

At the same time, all screen activity logs - collected by the cross-platform APIs (pyautogui, WinAPI, Quartz) - go to the Screen Activity Analyzer (SAA). It watches active window titles, tab switching frequency, and mouse/keyboard events density and dwell time in different categories of applications and learns the behavioral patterns with LSTM networks to identify patterns such as educational app - social media - idle. The screen Focus Score (SFS) is the measure of the screen utilization:

$$\boldsymbol{textSFS = 1 - (Idle\,Time/Total\,SessionTimeElapsed)}$$

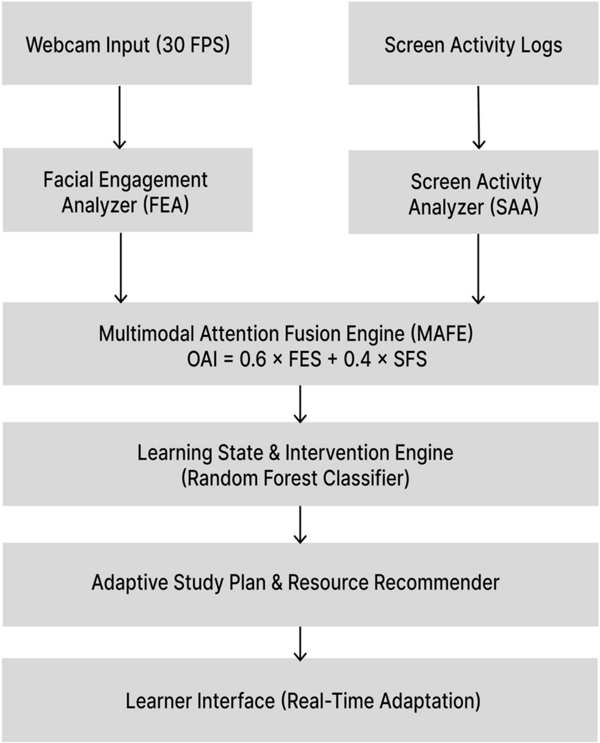
Webcam Input (30 FPS)
Screen Activity Logs
Facial Engagement Analyzer (FEA)
Screen Activity Analyzer (SAA)
Multimodal Attention Fusion Engine (MAFE)
OAI = 0.6 × FES + 0.4 × SFS
Learning State & Intervention Engine (Random Forest Classifier)
Adaptive Study Plan & Resource Recommender
Learner Interface (Real-Time Adaptation)

Fig.2 Proposed A^3ITS Architecture

Idle durations are the sum total of periods during which the input events are absent in the span of more than 5 consecutive seconds and it offers an intuitive proxy of on-task behavior that is resistant to different session durations.

3.1.3. Cognitive State Classification Layer: +Decision Logic: The Learning State & Intervention Engine (LSIE) takes OAI time-series at a sliding window of 1-minutes and derives statistical features (mean, variance, linear trend slope) on which Random Forest classification takes four discrete cognitive states, with each state triggering domain-appropriate responding:

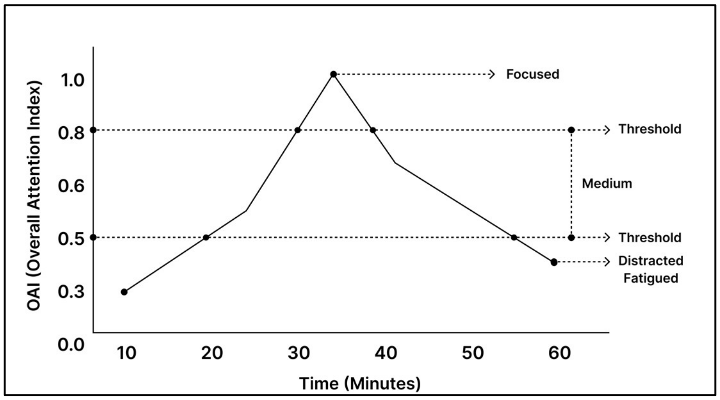

Fig.3 Attention Index (OAI) vs Time with Cognitive States

+Focused State (OAI > 0.8 and rising): As a result of profound understanding, the pace of the system, the introduction of extension exercises, or prerequisites.

+Medium Engagement (OAI 0.5- 0.8): Standard learning conditions; consistent normal instructional, but with selective micro level assessments.

+Distracted State (OAI < 0.5 by oscillation): Activates instant behavioral nudges soft TTS prompts ("We need to focus on the screen fast; would you mind doing 30 seconds of breathing exercises), content gamification or 30 seconds of breathing exercises.

+Overload State (OAI less than 0.5 and a downward trend): Triggers recovery programs such as five-minute periodic pauses, simplification of the content (with Flesch Kincaid grade level two) or refreshing multimedia breaks.

It is a threshold-based logic with session history added, and represents the principles of just-in-time pedagogy that was shown to reduce dropouts by 28% in meta-analyses.

3.1.4. Adaptive Output Layer:

+Customized Instructional Delivery:

ASPRR is an operationalization of state decisions that is based on hybrid content-user filtering. Sentence BERT embedding common is obtained with cosine similarity on educational video transcripts as a result of learner queries:

textResource Similarity = Cosine Similarity

(Query Embedding, Candidate Video Embedding)

State controlled by means of regulating it outputs: overload simplifies it; focus opens up advanced directions. The Chatbot YouTube Data API integration exposes the top-3 most similar videos that have overlaid engagement predictions, and the dynamic study plans reorganize the sequences in a study, prerequisites, and when to assess them. Difficulty calibration uses readability measures that guarantee the optimal cognitive load of cognitive load.

3.2 Datasets that are required to develop A3ITS:

The training of A3ITS needs multimodal datasets in the facial engagement cues and patterns of screens activity. There are ready-made benchmarks in public datasets, which were complemented with screen logs.pmc.ncbi.nlm.nih+2.

+ Facial Engagement Datasets:

Student Dataset (ICCVW 2021): Front camera annotations of videos of college students solving math problems, between the states of engage (focus on the screen/paper) and wander. Perfect FEA gaze/pose.openaccess.thecvf training.

Students Attention Detection Dataset (Mendeley): 4,000 records and are detected on the facial level, pose estimation (head orientation, rotation), hand tracking, and phone distraction. Directly supports FES.

DAiSEE Dataset: Videos of students categorized by degree of engagement (4-class: very low to highly engaged) by facial expression, eye closure, head movement.

+Screen Activity Datasets:

StudentLife Dataset (Dartmouth 2013): Detailed information of 48 students during 10 weeks, such as application use, interaction time, academic outcomes relationships.

iRemix Online Learning Logs: 88k+ student interaction logs (clickstreams, session times) of social learning platform, coded into 21st-century skills engagement learning platforms.

+**SCBehavior Dataset:** contains 1,346 images of classrooms that have been labeled with behavior (read, write and discuss); access through screen by use of proxy window using simulation. GitHub.

+Personalized Collection to drive Multimodal Fusion:

Synchronized webcam and screen logs of 50+ student volunteers at 30-min MOOC sessions using pyautogui to get window titles, idle time, switching of tabs. Label OAI ground truth through post session surveys.

3.3 Preprocessing Techniques

Preprocessing is very important in making sure that the data it utilizes in the calculation of FES, SFS, and OAI is clean, consistent, and in time. The system is based on a number of modalities which makes it at this stage highly important to handle its management with a great care so that it enhances the reliability of the downstream model.

+Processing facial data (MediaPipe Pipeline)

Videos of faces are initially sampled at 15-30 fps to trade-off computation flexibility with time response. MTCNN or MediaPipe will be used to detect the faces, and dense face landmarks will be extracted.

Per frame, 468 landmarks of faces are extracted and this can allow the calculation of: view vectors and orientation of the monitoring head (Euler angles such as the yaw, pitch and roll).

Blank interval of the eye blinking, defined through Drowsiness Aspect Ratio (DAR) threshold.

+OpenFace Facial Action Units (AUs)

All facial features are normalized to sigmoid scale in order to provide the feature consistency between the subjects and sessions. Noise that is introduced by the camera jitter or micro-movements over time, is eliminated by Exponentially Weighted Moving Average (EWMA) smoothing with = 0.7.

Data augmentation to enhance robustness and generalization is also implemented (horizontal flipping, brightness and contrast jitter plus -20, and -20) to replicate authentic lighting and camera fluxes of the real world.

+Processing of Screen Activity Log

Pyautogui logs are gathered to provide user interaction data, such as timestamps, name of the active window, mouse location, keyboard input and time idle.

A number of higher-level features are then obtained out of these raw logs:

The calculation of idle ratio is made at 30-sec rolling windows.

Classification based on application category (educational, social or idle) by using regex-based window title matching.

Tab-switching rate, in which a switch more than three switches/min considered a version of distraction signal.

In the case of sequence modeling, interaction windows are represented as tokenized and as LSTM-compatible inputs. The gaps in the logs which result in missing values are treated by forward-fill imputation to maintain the time series.

+Multimodal Synchronization and Fusion

To allow the multimodal learning to take place effectively, timestamps on the streams of facial and screen interaction can be synchronized with NTP on the synchronization of streams. Frames that are missing are reconstructed with the help of linear interpolation.

To be useful in fusion, FES and SFS attributes are organized into (T x 2) matrices, which is open to time-aligned analysis. The 60-second sliding window is used to compute the long-period engagement trends to be used in calculating the OAI.

The last data will be divided into 70 training, 15 validation and 15 testing where stratification is done based on the engagement tags to guarantee equal representation.

+Checks on Data quality and consistency

Sessions that are too idle are also rejected with the removal of the samples whose idle duration is more than 20 percent. In order to deal with the issue of class imbalance, SMOTE is used as required.

Through cleaning and synchronization, the dataset would comprise about 10,000 high-quality, multimodal samples, which are end to end trainable.

4. Theoretical Framework

The proponent A3ITS (Attention-Aware Adaptive AI Tutoring System) is designed based on various grounds of theoretical differentiation that overlap with the fields of cognitive psychology, affective computing, and adaptive learning paradigms (Worsley, 2014). This framework brings together both learning theories and computational modelling skills in an attempt to rationalize the use of multimodal attention and adaptive intelligence in customized e-learning systems.

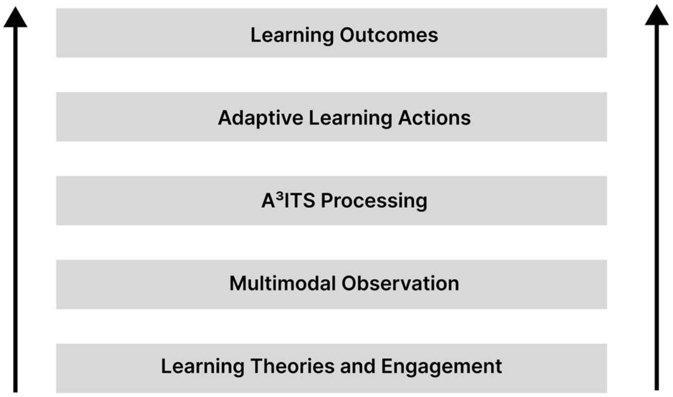

Fig.4 A^{3}ITS Conceptual Flow

4.1 Constructivist Learning theory and Cognitive Learning theory:

To the cognitive learning theory, the acquisition of knowledge is directed by and through active mental engagements, attention, and control of working memory. Constructivist standpoints also highlight the fact that learners construct knowledge by using interactive and contextual learning. A3ITS realizes these principles in its constant evaluation of learner attention and attention using the Facial Engagement Analyzer (FEA) and the Screen Activity Analyzer (SAA) by ensuring that instructional adaptation is depending on the cognitive readiness of the learner and accordingly the surrounding engagement (Pekrun, 2006).

4.2 Attention and Cognitive Load Theories:

Attention theory holds that, to ensure effective understanding and solidarity of memories, the constant capacity to concentration of attention is necessary. Cognitive Load Theory (CLT) which states that learning efficiency is hindered by heavy over mind or distraction. A3ITS quantifies behavioral and visual attention (Overall Attention Index (OAI)) to model the concentration of learners and dynamically adjust the instructional complexity reducing overload by responding to learners with timely interventions and changing the learning pace based on the Learning State and Intervention Engine (LSIE) (Wu et al., 2021).

4.3 Affective and Engagement Model

Affective Tutoring studies affirm the emotional aspect of learning where motivation, frustrations and fatigue play a major role in determining learning. The next step in the development of emotion-aware instructional systems is A3ITS, but it is a step, as the emphasis is put on practical behavioral responses rather than on emotions, to combine both affective and attentional signals. This is in line with the Control-Value Theory of achievement emotions whereby the emotion of learner's intermediates between cognitive control and perceived task value (Pekrun, 2006; Wu et al., 2021).

4.4 Personalization and Adaptive Learning Paradigms:

Informed by Intelligent Tutoring Systems (ITS) and Recommender Systems, adaptive learning models are used to support individualized instruction using learner profile and performance information. Leaving behind the rule-based approach to personalization, A3ITS proposes a real-time, multimodal adaptation layer, which is represented by the Adaptive Study Plan and Resource Recommender (ASPRR). This model is supported by Reinforcement Learning and feedback based pedagogical mechanisms that optimally learn the sequence of content on a continued system of the learner feedback response (Wu et al., 2021).

4.5 Multi-modal Learning Analytics Framework:

Based on the multimodal learning analytics (MMLA), A3ITS combines mixes of heterogeneous data streams, such as facial expressions, gaze, and on-screen actions, in inference of patterns of engagement. The Multimodal Attention Fusion Engine (MAFE) includes the fusion mechanism that encompasses the principles of data-driven learning analytics and sensor-informed modeling, which represent a part of overall cognition and behavior of learners in the digital environment.

This theoretical basis makes A3ITS an interdisciplinary framework of evidence-based application of cognitive, affective, behavioral, and adaptive theories of learning to facilitate contextually sensitive and continuous educational personalization.

5. Comparative Analysis

Dimension	Traditional ITS	Affect-Aware ITS	Attention-Aware Systems	Recommender-Based Learning Systems	Proposed A3ITS
Primary Data Sources	Quiz responses, problem-	Performance data + facial,	Gaze data, screen focus,	Clickstream data, usage history,	Quiz data + gaze, facial cues, posture, screen activity, interaction logs

	solving traces, dialogue logs	voice, posture cues	interaction logs	learner profiles	
Core Modeling Focus	Cognitive mastery and knowledge tracing	Learner affects and emotional state	Visual attention and on/off-screen focus	Content–learner similarity and preferences	Joint modeling of attention, engagement, affect, and learning state
Common Modeling Approaches	Rule-based systems, BKT, Bayesian networks, RL	Multimodal feature extraction, fusion, affect classification	Gaze tracking, attention heuristics	Collaborative filtering, content-based recommendation	Multimodal fusion (FES, SFS, OAI), state inference, adaptive policies
Granularity of Adaptation	Task selection, hinting, step-level feedback	Emotion-conditioned hints, encouragement, difficulty tuning	Alerts or basic content pauses	Content recommendation at session or module level	Dynamic pacing, content type, difficulty, intervention timing
Temporality of Personalization	Session-based, performance-driven updates	Near real-time affect detection with limited temporal fusion	Mostly real-time attention signals	Periodic or offline personalization	Continuous real-time + long-term learner state modeling
Attention Modeling	Implicit (errors, response	Implicit via multimodal affect	Explicit via gaze and focus	Indirect via engagement logs	Explicit multimodal attention modeling

	time proxies)	signals	tracking		

6. Discussion:

The relative comparison of classical ITS, affective tutoring systems, attention-aware architectures and recommender-based learning platforms reveal that a strong trend towards more adaptive and data driven personalization is emerging but also reveal constraints that exist in the structure. Viewed traditional ITSs are quite effective at modeling cognitive mastery providing step-by-step instructions but are mostly restricted to explicit performance feedback and domain-limited interaction forms or channels, which limit deep learning, learner autonomy, and robust engagement monitoring. Affective ITS also extend this perspective to include emotion recognition, but their interventions are rather emotion-focused, and they often pay little or no attention to observable behavioral signs of distraction or off-task activity in a digital study session.

Focus now on attention aware systems and multimodal learning analytics studies - imply that gaze, head pose and other such signals have proved useful in on task attention, but most current systems remain at the detection stage and does not provide significant curricular study or long-term modeling of learner conditions. On the same note, educational recommender systems have also facilitated the optimization of learning trajectories based on historical input data on interactions, although they usually execute in batch or other passive update modes and seldom process real-time changes in fatigue, overload or disengagement during a session. All these gaps mainly point towards the inadequacy of single-source or single-timescale models in the complexity of learner behavior in contemporary online

In this respect, the presented A3ITS framework adds an integrated multimodal view supporting the simultaneous assessment of visual (through the Facial Engagement Analyzer) and behavioral (through the Screen Activity Analyzer) engagement to maintain an estimate of the state of a persistent learner. Combining these signals into an Overall Attention Index and control Hen engine Learning State and Intervention Engine, the system real-time determination of focused, distracted, overloaded and states and mixes each grouping with any one of a set of responses (focus prompt, micro-break, simplification of the content). Such a design conforms to the recent ITS and AI-in-education literature calls to models that are adaptive as well as context-aware and responsive to context-noisy time-varying learner behavior. Adapt. Resource Recommender and customize study plans and order sequencing of resources based on changing engagement profiles to achieve personalization that is not just temporally personalized, but is adaptable in the real-time, session-by-session.

Simultaneously, A3ITS has its own challenges and design trade-offs which should be noted. The constant observation of the webcams and screen violate privacy, transparency, and consent, particularly concerning the formal aspect of the educational environment where the strict rules of data governance are observed. Strong fusion weight, threshold, and state label calibration is also not trivial: poor attention or overload classification might give out of context interventions and frustration to the learner. Moreover, the actual reviews of AI-based ITS also focus on the problem of generalizability and equity, indicating that models tend to become poorer when applied in new demographics, institutions, or devices. The next generation of A3ITS should thus include evaluation considerate of fairness, monitored levels under learner and instructor control and participatory and inclusive design with learners and instructors to be assured that the system positively influences learning but it does not increase biasness and other negative effects.

7. Future Work and Conclusion

The presented A3ITS (Attention-Aware Adaptive AI Tutoring System) develops a novel multimodal architecture that combines facial engagement analysis (FES) and screen activity monitoring (SFS) into an Overall Attention Index (OAI = 0.6xFES + 0.4xSFS), which allows detecting cognitive states and applying interventions in real time. A3ITS was the first to combine behavioral awareness and dynamically changing the study plan, unlike traditional ITS, gaze-only, or stationary recommenders, directly tackling MOOC dropout issues using non-invasive, edge-deployable intelligence.

Its future directions involve more multi modal sensing, adaptive fusion weights by reinforcement learning, massive pilots on SWAY AM/NEP systems, design of privacy that is offer able with interpretable explanations, and teacher-in-loop dashboards offering classroom interventions. Such guidelines will confirm influence on learning outcomes as well as ethical implementation.

A3ITS takes intelligent tutoring in the direction of situationally aware systems that respond not only to the gaps in the knowledge, but to the moment-to-moment realities of learner attention and struggle, to offer new opportunities to the technically robust, peda sound, ethically responsible AI.

References

- Acosta, H., et al. (2024). Multimodal learning analytics for predicting student engagement and collaborative behaviors.
- AI in education: Personalized learning and intelligent tutoring systems. (2025).
- Alkhatlan, A., & Kalita, J. (2019). Intelligent tutoring systems: A comprehensive historical survey. *International Journal of Computer Applications*.
- Artino, A. R. (2012). Control-value theory: Using achievement emotions to improve understanding of motivation, learning, and performance in medical education. *Medical Education, 46*(11), 1066–1078.
- Echeverria, V., et al. (2024). Exploring design considerations for multimodal learning analytics systems: An interview study. *Frontiers in Education*.
- Educational recommender systems: A systematic literature review. (2023). *BCE*.
- Giannakos, M., Sharma, K., Pappas, I. O., Kostakos, V., & Velloso, E. (2022). Multimodal data fusion in learning analytics: A systematic review. *Sensors, 22*(3), 1–30.
- Gitnux(**2026).** *E-learning statistics and market size report.* https://gitnux.org/e-learning-statistics/
- IJETT. (2023). A systematic literature review on the implications of recommender systems in education: Emphasizing ethical, explainable, and cross-regionally deployable ERS.
- J. V. P., et al. (2025). Multimodal learning analytics (MMLA) in education.
- Klašnja-Milićević, A., et al. (2022). A systematic literature review on educational recommender systems for teaching and learning: Research trends, limitations and opportunities. *Education and Information Technologies*.
- Létourneau, R., et al. (2025a). A systematic review of AI-driven intelligent tutoring systems (ITS) in K–12 education. *npj Science of Learning*.
- Létourneau, R., et al. (2025b). Systematic review of AI-driven ITS: Ethical implications and the need for investigating AI-in-teaching ethics.
- Moga, H., et al. (2014). Affective tutoring system based on extended control-value theory. *Procedia – Social and Behavioral Sciences, 141*, 518–523.
- Ochoa, X. (2017). Multimodal learning analytics. In *Handbook of learning analytics*.
- Ochoa, X. (2022). Multimodal learning analytics: Rationale, process, examples. In *Handbook of learning analytics* (2nd ed.). Society for Learning Analytics Research.
- Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice.

- Wu, C.-H., et al. (2021). Applying control-value theory and unified theory of acceptance and use of technology to explore learning emotions in online learning.
- Worsley, M. (2014). Multimodal learning analytics: Enabling the future of learning through multimodal data analysis and interfaces. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*.