AI-assisted modeling of attention quantifies engagement and predicts cognitive improvement in older adults

Yang Merik Liu[1], Adam Turnbull[1], Meishan Ai[1], Mia Anthony[1], Kathi L. Heffner[2], Cristiano Tapparello[3], Ehsan Adeli[1], Feng Vankee-Lin[1,4]

Affiliation:

1. *Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, CA*
2. *School of Nursing, University of Rochester, Rochester, NY*
3. *Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY*
4. *US Department of Veterans Affairs (VA) Palo Alto Health Care System, Palo Alto, CA*

Corresponding author: F. Vankee Lin, 1070 Arastradero road, Palo Alto, CA 94304 (email: vankee_lin@stanford.edu)

Acknowledgement: Authors claimed no conflict of interest.

Abstract

Background Cognitive training aims to prevent or slow cognitive decline in older adults, but outcomes vary widely. Engagement, describing how individuals allocate cognitive, affective, and physiological resources, is critical to training benefits, yet behavioral metrics lack real-time modeling of attention and do not reliably predict outcomes. We developed and validated a multimodal, AI-assisted biomarker that quantifies attentional states during computerized cognitive training and predicts cognitive improvements.

Methods We designed the Attentional Index using Digital measures (AID), leveraging a video-based facial expression encoder (pretrained on 38,935 videos), an ECG-based autonomic encoder (pretrained on 123,998 ECG samples), and a temporal fusion module. Using two processing speed/attention studies in older adults (> 65 years) with mild cognitive impairment, AID was trained and evaluated in BREATHE (n=50; ~300 hours from 368 sessions) and validated in FACE (n=20; ~150 hours from 219 sessions). Model training targeted session-level change in self-reported fatigue. Clinical validation tested relationships between AID scores and (1) behavioral attention, (2) cognitive outcomes, and (3) neural correlates.

Findings AID accurately detected engagement changes (BREATHE: accuracy 0.82, F1=0.81; FACE: accuracy 0.73, F1=0.74), outperforming unimodal models. AID scores were unrelated to session, task type, or demographics. In BREATHE, session-level AID scores significantly predicted executive function improvement (session×AID: Wald χ^2=7.85, p=0.005), whereas reaction time variability did not. Lower AID intercepts (B=-0.07±0.03, p=0.043) and steeper slopes (B=0.31±0.15, p=0.046) were associated with greater improvements. Post hoc analyses identified two engagement profiles linked to better attention: one characterized by low-RMSSD and focused periocular activation, and the other defined by coherent alignment between low-RMSSD and facial expression patterns.

Interpretation AID provides a reliable digital biomarker of effective engagement and predicts cognitive improvement beyond behavioral metrics. By capturing facial-autonomic dynamics of attention, AID offers a foundation for closed-loop cognitive intervention design.

Funding NIH AG081723, NR015452, and AG084471; Stanford HAI seed funding.

Research in context

Evidence before this study

We searched PubMed, Google Scholar, Web of Science, IEEE Xplore, and reference lists of relevant reviews for studies published in English before March 1, 2025. Search terms included combinations of "cognitive training," "engagement," "attention," "older adults," "mild cognitive impairment," "heart rate variability," "facial expression," "psychophysiology," "multimodal," "machine learning," *and* "digital biomarkers." We included studies involving older adults, computerized cognitive training (CCT), and behavioral markers of engagement/attention. We excluded studies that did not involve aging populations, did not report attention or engagement measures, or lack objective psychophysiological or behavioral signals.

Across the evidence base, behavioral performance metrics (reaction time, accuracy) showed inconsistent associations with training-related cognitive outcomes and lacked the capacity to capture attentional dynamics. Psychophysiological markers reflected arousal or effort but were rarely linked to cognitive transfer effects. Importantly, no study identified through our search evaluated a real-time, multimodal framework combining facial and autonomic signals to quantify attention during CCT, nor did any study provide external validation across independent cohorts. The quality of existing evidence was moderate, with common limitations including small samples, limited generalizability, and high risk of bias due to variations in self-reports.

Added value of this study

This study establishes, for the first time, a multimodal, AI-assisted biomarker that quantifies attentional states during CCT by integrating facial and autonomic features. The AID framework was comprehensively validated, demonstrating reliable performance across two independent aging cohorts and robustness across tasks and sessions, and more accurately predicted cognitive improvement than conventional behavioral metrics. Our findings introduce a clinically interpretable engagement biomarker that correlates with neural and psychophysiological signatures of attention, overcoming limitations of behavioral-only approaches.

Implications of all the available evidence

In summary, prior evidence and our findings suggest that multimodal measures integrating facial and autonomic signals may provide a more detailed description of effective engagement during cognitive training by modeling both the attentional availability and allocation. Such measures could eventually help refine non-pharmacological interventions for older adults at risk for cognitive decline and inform future research in personalized, closed-loop cognitive training design. However, although AID shows promise as an objective and generalizable indicator of attentional state, further validation in larger and more diverse samples is required. At this stage, AID should be regarded as a tool that contributes to understanding how attentional dynamics relate to training response.

Introduction

Cognitive interventions represent a promising frontline for preventing or slowing cognitive decline and brain aging in individuals at risk for Alzheimer's disease (AD, brief as "AD-risk group")[1]. Among these, computerized cognitive training (CCT), a suite of computer-based tasks designed to enhance specific cognitive functions, has shown potential to promote healthy brain aging due to its scalability, affordability, and safety profile[2], where CCT targeting processing speed/attention (PS/A) has shown the most promise[3]. However, evidence regarding the effectiveness of CCT remains mixed, in part due to substantial variability in adherence and engagement[4]. Particularly, there is a poor understanding of "effective engagement", or the best way for participants to invest their cognitive, affective, and physical energies in the intervention to encourage improvements in clinically meaningful improvements (i.e., transfer to at-risk cognitive domains)[5]. Existing markers of engagement/adherence, such as performance on training tasks or time/frequency/duration of training, are poor predictors of these outcomes[6]. There is a need for improved markers that can capture the complex dynamics involved in the engagement process, including balances between competing processes in the AD-risk group.

Signals capturing attention are particularly promising given the central role of attention in engagement[7]. Regulating attention requires the integration of top-down goals with bottom-up stimuli, which are both central to cognitive training outcomes[8] and the primary target for PS/A training[9]. The dorsal and ventral attention networks, respectively, have been proposed to coordinate these separable but complementary aspects of attention[8]. In the context of CCT, attention encompasses at least two dimensions related to the complex interplay between cognitive demand, engagement experience, and cognitive plasticity: (1) *attentional availability*, or the amount of attentional resources available to be allocated to the task; and (2) *attentional allocation*, or the current energy invested in the task in a given moment. The two dimensions emerge over different timescales, with availability following fluctuations in arousal that occur over tens of seconds of time or longer, while allocation changes moment-to-moment in response to task demands[10]. However, they also interact such that availability can change instantly in response to specific environmental cues, and allocation is also influenced by fluctuating arousal[11]. We propose that two conditions may be particularly relevant for encouraging the plasticity needed for effective engagement: (1) maximized availability of attentional resources; (2) maximized allocation of available resources or the coherence between availability and allocation.

Conventional tools struggle to capture these complex nonlinear interactions, especially in the AD-risk group that show increased variability and noise in attention[12]. Advances in artificial intelligence (AI), such as deep neural networks and temporal encoders, offer a transformative approach to capturing the dynamic and multimodal nature of attention during CCT[13]. Critically, these methods allow the extraction of critical signals from objective markers and do not require interruption of the task as is necessary for self-report measures, allowing an improved quantification of the attention process, rather than just the outcome reflected in task performance[14]. Specifically, parasympathetic nervous system activity, via ECG-derived signals, shows a suppression-and-rebound parasympathetic activity pattern during cognitive training: when a task is perceived as challenging, resources are made available to meet the task demand by reducing those available for parasympathetic regulation, with resources resuming upon adaptation to the task[15–17]. Meanwhile, temporal facial gestures, especially subtle periocular movements, have been shown to mirror moment-to-moment fluctuations in mental effort and

engagement during sustained cognitive tasks, reflecting the dynamic allocation of attention[18]. By fusing facial and psychophysiological signals, AI may capture critical information about the efficiency and sustainability of the attention process during CCT that can act as an index of effective engagement.

Here we proposed a novel marker of *Attentional Index using Digital measures (AID),* leveraging a vision transformer model (ViT-B/16[19]) for facial expression (FE) and a time-series modeling approach (TimesNet[20]) for autonomic dynamics to quantify real-time attention. Specifically, we developed, validated, and interpreted AID as an indicator of effective engagement across two similar intervention datasets (BREATHE and FACE: both performing CCT in AD-risk groups). We hypothesized that a multimodal approach integrating FE and ECG might more accurately reflect effective engagement than either modality alone.

Methods

Study design and participants

We used two independent datasets, i.e., BREATHE[21] and FACE, each comprising AD-risk groups who completed speed of processing training (SOPT) tasks over multiple CCT sessions with synchronized facial video and ECG recording (BREATHE: ~300 hours' data from 368 sessions; FACE: ~150 hours' data from 219 sessions; see **Figure 1A** for CCT protocol setup overview and **Table 1** for sample characteristics). Details of BREATHE and FACE protocols, participant eligibility, SOPT task paradigms, and data preprocessing procedures are described in **Online Materials**. Both study protocols were approved by institutional IRBs and participants' informed consents were obtained.

Across both datasets, self-reported fatigue was collected at the start and end of each session using a 6-item visual analogue scale[22] (0-5). From these ratings, we derived two distinct supervision signals: (1) the <u>session-level fatigue</u> was created by binarizing the mean fatigue score into *fatigue* ($\geqslant 2$) versus *minimal/no fatigue* (<2) for learning spatial representation of engagement; (2) we created the <u>session-level Δfatigue</u> (post-session – pre-session), with session showing an increase of $\geqslant 1$ point coded as *fatigue increase* and all others as *no increase*, for learning temporal variation of engagement. Moreover, an additional self-reported tiredness rating (0-5) was collected every 2-minute blocks during each session in FACE; these block-level ratings were converted similarly, i.e., *fatigue* (tiredness$\geqslant 2$) versus *minimal/no fatigue* (tiredness <2), and served as the <u>block-level fatigue</u> to support model finetuning. These three supervision signals were used in strictly separated training pathways without any leakage.

AID framework development

The AID framework was developed on the BREATHE dataset to predict <u>session-level Δfatigue</u>, given the relevance of fatigue increases as an indicator of inefficient or unsustainable attention in AD-risk groups[23,24]. It consists of two modeling stages: (1) modality-specific encoders for FE and ECG data, and (2) a multimodal fusion module integrating both modalities over time to predict fatigue change. **Figure 1B** illustrate the data partitioning strategies for BREATHE and FACE, respectively.

<u>*Stage 1: Modality-Specific Encoders (see **'Stg-1' in Figure 1C**)*</u>

FE encoder: We employed a Vision Transformer (ViT-B/16)[19], pretrained on the FERV39k Multi-Scene Dataset[25] (incl. 38,935 video samples with up to 1 million facial expression images), to encode frame-level FE features. Each 160×160 cropped face image was divided into 16×16 patches, flattened, and linearly projected into patch features, augmented with positional encodings. These features were passed through multiple transformer layers to capture global spatial dependencies across facial regions, enabling detection of subtle cues such as brow tension, gaze shifts, or micro-expressions indicative of attentional allocation and fatigue (see **Figure 1C left**). Frame-level outputs were temporally aggregated over 5-second clips to form dynamic features representing FE patterns across sessions.

ECG encoder: ECG signals were encoded using TimesNet[20], a temporal foundation model optimized for time-series data, pretrained on ECG Heartbeat Categorization Dataset[26] (incl. 123,998 ECG samples from 341 subjects). After preprocessing (artifact removal, R-peak detection, 5-second segmentation), ECG sequences were fed into TimesNet's multi-scale temporal convolutional architecture. This design captures both short-term variability and long-range patterns in autonomic activity. The model learns interpretable representations of attentional flow, such as transient parasympathetic suppression or HRV inflections, aligned with task transitions and fatigue accumulation (see **Figure 1C right**). Each 5-second ECG segment was converted into a feature that captures the evolving psychophysiological state during cognitive engagement.

During Stage 1, each encoder (FE and ECG) was trained to predict *fatigue* vs. *minimal/no fatigue*. The first and last 5% data of each session in BREATHE were annotated using pre-/post-session-level fatigue, while the middle 90% remained unlabeled and contributed to representation learning. A binary cross-entropy loss was then applied to supervise the encoder training based on these binarized labels (see **'Label' in Figure 1C**).

*Stage 2: Multimodal Fusion Module (see **'Stg-2' in Figure 1D**)*

To integrate behavioral and psychophysiological signals into a unified representation of attention, we employed a multimodal fusion framework that combines video-based FE and ECG-derived HRV after encoder training. For each 30-second window (not overlapped), synchronized 5-second segments of video and ECG data were first encoded independently using the corresponding encoders, producing aligned sequences of modality-specific features. These sequences were concatenated at each timestep and input into a shared Long Short-Term Memory (LSTM) network[27], which captures temporal dependencies across both modalities. The LSTM learns to model both within- and cross-modality dynamics, such as how shifts in FEs co-occur with regulation during attentional transitions. Hidden states from the LSTM were aggregated using attention-based temporal pooling to generate a session-level multimodal feature. These multimodal features were passed to a downstream classifier for predicting session-level fatigue changes, enabling robust, real-time engagement inference (see **Figure 1D**).

In Stage 2, the full AID framework was trained to predict session-level Δfatigue, i.e., *fatigue increase* vs. *no increase*. Binary cross-entropy loss was again applied to supervise classification, with optimization via the Adam optimizer and early stopping based on validation loss. Pretrained encoder weights were initially frozen during fusion training and later unfrozen for end-to-end fine-tuning.

Model training followed a cross-subject protocol, where *Stage 1* encoder training and *Stage 2* multimodal fusion training were both conducted using the same subject-level partitions, ensuring

that no participant's data contributed to both training and evaluation in either stage. Model parameters were selected based on the best average F1 score. All experiments were seeded to ensure consistent results across runs. Detailed model training configurations are described in **Online Materials**.

AID framework evaluation

To further examine the developed AID framework, we evaluated it on the FACE dataset with a smaller and independent sample. The block-level fatigue in FACE was used to provided fine-grained supervision for representation learning. Moreover, we assessed cross-cohort transferability by applying the pre-trained model on BREATHE and fine-tuning it end-to-end using 50% of the FACE data. The AID performance was then evaluated on the remaining 50% of FACE. Fine-tuning included both encoder and fusion components initialized with pretrained BREATHE weights. All components of the AID framework, including modality-specific encoders and the multimodal fusion module, were implemented same as in BREATHE, ensuring fair evaluation and no data leakage.

AID framework clinical validation

AID scores were computed for each 30-s block as the mean LSTM output from concatenated facial-expression and ECG features as the basis for the following analyses: (1) To evaluate the temporal dynamics of engagement, both over-session and between-task, we calculated std-RT over trial-based RT to accurate response during SOPT tasks for each CCT session across two datasets, where higher values indicated worse attention; (2) To examine the association with clinical outcomes, EF was assessed at baseline and post-intervention using the NIH EXAMINER battery[28] in BREATHE, where ΔEF was calculated as the difference with higher scores indicating greater improvement; (3) To further investigate the neural correlations, in BREATHE, resting-state fMRI at baseline quantified functional connectivity within five canonical attention networks (dorsal attention, ventral/salience, limbic, default mode, and control), following Yeo's seven-network parcellation (see **Online Materials**).

AID framework post hoc interpretation

To analyze within-session engagement dynamics in BREATHE, we aligned 30-s windows of FE and ECG features with their corresponding AID scores. Parasympathetic activity was indexed using RMSSD derived from 30-s ECG segments. For feature-space visualization, we selected the top 10% of segments with each session based on the LSTM's temporal-attention weights[27], which represent each segment's relative contribution to predicting Δfatigue. These top-ranked segments were assumed to represent periods of peak attention and effective engagement. Multimodal features (concatenated FE and ECG features) from these segments were pooled across session and subjects. Using RMSSD from these segments as an interpretable proxy of autonomic availability, we performed k-means clustering to identify autonomic patterns of engagement and then characterized the associated FE patterns within each cluster, rather than clustering directly in the high-dimensional multimodal feature space. The optimal clustering solution was determined by elbow and silhouette criteria. Analyses of clinical validation and interpretation were primarily conducted using generalized linear model for cross-sectional data or generalized estimating equation for timeseries data. Detailed processing pipelines, statistical equations, and RMSSD calculation, and clustering results are provided in the **Online Materials**.

Results

Results of AID framework evaluation

Table 2 summarizes AID performance for predicting within-session Δfatigue across modalities, as well as cross-dataset validation results. Importantly, in BREATHE, Face+ECG significantly outperformed both Face and ECG alone ($p < .05$). No significant differences were observed in the FACE dataset. The BREATHE pretrained multimodal model showed moderate generalization in the FACE validation cohort, highlighting the complementary roles of FE and ECG in measuring engagement.

Results of AID framework clinical validation

Over-session engagement patterns: In BREATHE, using GEE model taking the session-based AID score as outcome, and session and intervention group assignment (brief as "group" afterward) as predictors (*Equation 1.1*), there was no significant difference in AID score over session (Wald's χ^2=1.44, p=0.23), suggesting that the engagement level was not influenced by sessions (see **Figure 2A**). In BREATHE, there was no difference in AID scores between MCI and HC group (Wald's χ^2=0.23, p=0.88, see **Figure 2A**) when using the GEE model taking AID scores as outcome, and session, group, and cognitive status as predictors (*Equation 1.2*).

Between-task engagement patterns: In FACE, using GEE model, taking AID scores as outcome, and session and task type as predictors (*Equation 2.1*), there was no significant difference in AID scores between tasks (Wald's χ^2=5.24, p=0.16), suggesting that the engagement level was not influenced by task type (see **Figure 2B**).

Association with neural correlates of attention: When examining the relationship between AID scores and baseline brain networks potentially related to attention (i.e., dorsal attention, salience/ventral attention, limbic, default mode, and control) using GEE model taking AID score as outcomes, and session, group and five networks as predictors (*Equation 1.3*), ventral attention (B=3.26, SE=1.60, Wald's χ^2=4.13, p=0.042) and limbic (B=-3.66, SE=1.78, Wald's χ^2=4.24, p=0.040) network significantly predicted session-based AID score (see **Figure 2C**); note that when applying false discovery rate correction, significant results did not survive. When applying cognitive status (i.e., MCI or HC) to *Equation 1.3* as an additional predictor, results remained similar. Also, there was no difference between demographics (i.e., age, sex, and education) and AID scores when using *Equation 1.4*.

Association with CCT outcome and behavioral metrics of attention: In BREATHE, we used ΔEF as outcome by examining two models: (base model) group, session, session-based std-RT, and session×std-RT (*Equation 1.5a*); and (expanded model) base model, session-based AID scores, and session×AID score (*Equation 1.5b*). There was no relationship between session-based std-RT and ΔEF (session×std-RT term: Wald's χ^2<0.01, p=0.99); however, there was a significant relationship between session-based AID scores and ΔEF (session×AID score: Wald's χ^2=7.85, p=0.005), highlighting AID as a clinically meaningful engagement biomarker. For interpretation purpose, we calculated the intercept and the linear slope of the AID score across session and plotted their relationships with ΔEF controlling for group. Smaller intercept (B=-0.07, SE=0.03, t=-2.07, p=0.043) and greater slope (B=0.31, SE=0.15, t=2.04, p=0.046) of AID score over sessions were related to greater ΔEF (see **Figure 2D**). In FACE, using a GEE model with session-based std-RT as outcome and session, task type, and AID scores as predictors (*Equation*

2.2), AID scores were significantly associated with std-RT (B=0.10, SE=0.03, Wald's χ^2=10.93, p<0.001), suggesting that smaller AID scores corresponded to better attention (see **Figure 2E**).

Results of AID framework post hoc interpretation

K-means clustering of aggregated RMSSD values in BREATHE identified two autonomic engagement patterns: (1) *Low-RMSSD cluster*, characterized by lower parasympathetic activity and greater periocular activation, which is consistent with focused engagement; (2) *High-RMSSD cluster*, showing elevated RMSSD and greater corrugator/cheek activation, which is consistent with disengagement or fatigue. Importantly, participants in the *High-RMSSD cluster* exhibited significantly higher std-RT, indicating poorer attention. **Figure 3A** displays the results on k-means clustering of the top-ranked 30-second aggregated RMSSD values and corresponding FE patterns.

Then, for each FACE session, we calculated the proportion of 30-second segments whose RMSSD values or FE patterns resembled either of the two identified clusters based on the distance to the nearest cluster centroid. To quantify multimodal alignment, we assessed whether RMSSD-based and FE-based cluster assignments converged within a session. Sessions were classified as *coherent* when more than 30% of segments were concurrently assigned to the same cluster (either *Low-RMSSD* or *High-RMSSD*) across both modalities; all other sessions were labelled *incoherent*. Analyses taking std-RT as outcome, and sessions, task type, and proportion score of RMSSD resembling either cluster as predictors (*Equation 2.3a*), showed that FACE sessions resembling *Low-RMSSD* alike patterns tended to have lower std-RT (B=-0.09, SE=0.05, Wald's χ^2=3.39, p=0.066), whereas *High-RMSSD* alike sessions tended to have higher std-RT (B=0.11, SE=0.14, Wald's χ^2=3.80, p=0.051; **Figure 3B**). Sessions dominated by coherent multimodal patterns had significantly lower std-RT compared with incoherent sessions (B=-0.14, SE=0.06, Wald's χ^2=5.43, p=0.020; **Figure 3C**), taking std-RT as outcome, and session, task type, and coherence vs. incoherence variable as predictors (*Equation 2.3b*). Together, these results demonstrate that multimodal engagement profiles learned in BREATHE, indexed jointly by HRV (availability) and FE patterns (allocation), replicate in an independent dataset and consistently track attentional dynamics. Incoherence between autonomic and facial signals further marks poorer attention, supporting AID's validity as an indicator of attention rather than a general fatigue detector.

Discussion

By integrating video-derived FE patterns and ECG-derived HRV patterns, the AID framework quantifies real-time attentional states in 30-second intervals associated with fatigue-related disengagement and preserved attentional allocation. These moment-to-moment attentional states also predicted ΔEF. Importantly, AID proved robust against individual differences in demographic factors (age, sex, education) and baseline cognitive status (HC vs. MCI), demonstrating its generalizability. Our findings further validate the availability-allocation model of attention, emphasizing both availability of resources (via HRV) and the coherence between availability and allocation (via coupled FE and HRV patterns). Two attentional patterns emerged consistently: a low-RMSSD pattern with focused periorbital activity, and a high-RMSSD pattern with heightened corrugator and cheek activation. These patterns corresponded to high and low attentional states. Overall, AID provides a robust and interpretable multimodal framework for capturing effective engagement during CCT.

Prior literature highlights the challenge in linking performance in CCT sessions to improvements in transfer effects, particularly outcomes not directly practiced during training[21]. Consistent with this, our BREATHE findings confirmed no relationship between session-based attentional metrics (std-RT) and post-intervention EF improvement. In contrast, session-based AID scores, which are related to both behavioral and neural signatures of attention, predicted ΔEF. Across datasets, lower AID scores indicated poorer attention. Moreover, in BREATHE, individuals who began with lower attention (lower intercept of over-session AID score) but whose engagement improved across sessions (slope) demonstrated greater EF enhancements, suggesting that participants with initially weaker attentional conditions may receive substantial benefit from interventions when their engagement improves over time.

We also observed no significant change in AID score across sessions when controlling for intervention group assignment in BREATHE. Because intervention groups remain blinded, we have not yet analyzed if there were differences in AID scores between groups or if AID scores are linked differently to cognitive improvements between groups. Once group identities are revealed, and as future CCT protocols incorporate AID-enabled real-time FE and ECG monitoring, we anticipate improved prediction of transfer effects both at the individual and session levels.

We confirmed our earlier perspective that effective engagement is a multimodal construct[6]. The multimodal model consistently outperformed unimodal models in predicting Δfatigue across datasets. The availability-allocation model also requires considering multiple engagement scenarios: both high resource availability and strong coherence between availability and allocation contribute uniquely to individual-level and momentary-level attention (see **Figure 3D** for illustration).

In terms of specific HRV and FE patterns, high attentional capacity was associated with low RMSSD alongside FE patterns focusing on eye and periorbital regions. Although prior work in healthy younger adults often associate RMSSD, an index of PNS, with greater cognitive capacity. Our contradictory finding here aligns with emerging literature on older adults with cognitive impairment and our earlier work[17,29]. That is, when older adults engage intensively in cognitively demanding tasks, they may withdraw brain resources that regulate the PNS to focus on the task, leading to a suppression of PNS signals. Conversely, when individuals feel fatigued or disengaged, PNS signals rebound, or the brain resources are not directed to attend challenging tasks. These results were replicated across both datasets. Similarly, the identified FE patterns, eye/periorbital regions vs. corrugator/cheek regions, are consistent with known behavioral distinctions between cognitive and affective regulation[29,30].

Several limitations are worth discussing. Both datasets represent modest and demographically narrow samples; larger and more diverse cohorts are needed to validate generalizability. Next, while AID scores were predictive of EF change, relationships with other cognitive domains (e.g., memory) remain unexplored, so as AID's causal influence on cognitive improvement. Once blinding is lifted in BREATHE, and as future studies test real-time adaptive CCT paradigms informed by AID, these mechanistic questions can be examined more directly. Finally, given the salience of eye/periorbital activities, future work may refine AID's temporal granularity and incorporate richer oculomotor signals as pupillometry.

Overall, these results support the utility of AID as a sensitive and robust measure of engagement that generalizes across tasks, sessions, and participant samples, laying a foundation for

strengthening effective cognitive engagement by tailoring it to attentional control dynamics among older adults with AD risk.

References

1. Rebok GW, Gallo JJ, ThorpeJr. RJ. Advancing Alzheimer's Disease and Related Dementias Research: The Johns Hopkins Alzheimer's Disease Resource Center for Minority Aging Research. *J Aging Health*. 2025;37(3-4_suppl):3S-8S. doi:10.1177/08982643241308448

2. Hill NTM, Mowszowski L, Naismith SL, Chadwick VL, Valenzuela M, Lampit A. Computerized Cognitive Training in Older Adults With Mild Cognitive Impairment or Dementia: A Systematic Review and Meta-Analysis. *AJP*. 2017;174(4):329-340. doi:10.1176/appi.ajp.2016.16030360

3. Edwards JD, Xu H, Clark DO, Guey LT, Ross LA, Unverzagt FW. Speed of processing training results in lower risk of dementia. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*. 2017;3(4):603-611. doi:10.1016/j.trci.2017.09.002

4. Hill NTM, Mowszowski L, Naismith SL, Chadwick VL, Valenzuela M, Lampit A. Computerized Cognitive Training in Older Adults With Mild Cognitive Impairment or Dementia: A Systematic Review and Meta-Analysis. *American Journal of Psychiatry*. Published online April 1, 2017. doi:10.1176/appi.ajp.2016.16030360

5. Ball K, Berch DB, Helmers KF, et al. Effects of Cognitive Training Interventions With Older AdultsA Randomized Controlled Trial. *JAMA*. 2002;288(18):2271-2281. doi:10.1001/jama.288.18.2271

6. Guimarães AL, Lin FV, Panizzutti R, Turnbull A. Effective engagement in computerized cognitive training for older adults. *Ageing Research Reviews*. 2025;104:102650. doi:10.1016/j.arr.2024.102650

7. Nahum-Shani I, Shaw SD, Carpenter SM, Murphy SA, Yoon C. Engagement in digital interventions. *American Psychologist*. 2022;77(7):836-852. doi:10.1037/amp0000983

8. Hopfinger JB, Buonocore MH, Mangun GR. The neural mechanisms of top-down attentional control. *Nature neuroscience*. 2000;3(3):284-291.

9. Lin FV, Tao Y, Chen Q, et al. Processing speed and attention training modifies autonomic flexibility: A mechanistic intervention study. *NeuroImage*. 2020;213:116730. doi:10.1016/j.neuroimage.2020.116730

10. Raut RV, Snyder AZ, Mitra A, et al. Global waves synchronize the brain's functional systems with fluctuating arousal. *Science Advances*. 2021;7(30):eabf2709. doi:10.1126/sciadv.abf2709

11. Robbins TW. Arousal systems and attentional processes. *Biological Psychology*. 1997;45(1):57-71. doi:10.1016/S0301-0511(96)05222-2

12. Peters F, Ergis AM, Gauthier S, et al. Abnormal temporal dynamics of visual attention in Alzheimer's disease and in dementia with Lewy bodies. *Neurobiology of Aging*. 2012;33(5):1012.e1-1012.e10. doi:10.1016/j.neurobiolaging.2011.10.019

13. Hallion LS, Hsu KJ, Schleider JL. Cognitive training for mental health problems. *Nat Mental Health*. 2024;2(1):17-24. doi:10.1038/s44220-023-00185-y

14. Sugden N, Thomas ,Matt, and Kiernan M. A scoping review of the utility of self-report and informant-report prospective memory measures. *Neuropsychological Rehabilitation*. 2022;32(6):1230-1260. doi:10.1080/09602011.2021.1875851

15. Chen Q, Yang H, Rooks B, et al. Autonomic flexibility reflects learning and associated neuroplasticity in old age. *Human Brain Mapping*. 2020;41(13):3608-3619. doi:10.1002/hbm.25034

16. Lin F, Heffner KL, Ren P, Tadin D. A Role of the Parasympathetic Nervous System in Cognitive Training. *Current Alzheimer Research*. 2017;14(7):784-789. doi:10.2174/1567205014666170203095128

17. Lin FV, Heffner KL. Autonomic nervous system flexibility for understanding brain aging. *Ageing Research Reviews*. Published online 2023:102016.

18. Wang Y, Turnbull A, Xu Y, Heffner K, Lin FV, Adeli E. Vision-based estimation of fatigue and engagement in cognitive training sessions. *Artificial Intelligence in Medicine*. 2024;154:102923. doi:10.1016/j.artmed.2024.102923

19. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv*. Preprint posted online June 3, 2021. doi:10.48550/arXiv.2010.11929

20. Wu H, Hu T, Liu Y, Zhou H, Wang J, Long M. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. *arXiv*. Preprint posted online April 12, 2023. doi:10.48550/arXiv.2210.02186

21. Lin FV, Heffner K, Gevirtz R, Zhang Z, Tadin D, Porsteinsson A. Targeting autonomic flexibility to enhance cognitive training outcomes in older adults with mild cognitive impairment: study protocol for a randomized controlled trial. *Trials*. 2021;22(1):560. doi:10.1186/s13063-021-05530-z

22. Lee KA, Hicks G, Nino-Murcia G. Validity and reliability of a scale to assess fatigue. *Psychiatry Research*. 1991;36(3):291-298. doi:10.1016/0165-1781(91)90027-M

23. Lin FV. A Multi-Dimensional Model of Fatigue in Old Age: Implications for Brain Aging. *The American Journal of Geriatric Psychiatry*. 2023;31(2):152-161. doi:10.1016/j.jagp.2022.10.007

24. Kok A. Cognitive control, motivation and fatigue: A cognitive neuroscience perspective. *Brain and Cognition*. 2022;160:105880. doi:10.1016/j.bandc.2022.105880

25. Wang Y, Sun Y, Huang Y, et al. FERV39k: A Large-Scale Multi-Scene Dataset for Facial Expression Recognition in Videos. *arXiv:220309463 [cs]*. Published online March 20, 2022. Accessed March 25, 2022. http://arxiv.org/abs/2203.09463

26. Kachuee M, Fazeli S, Sarrafzadeh M. ECG Heartbeat Classification: A Deep Transferable Representation. In: *2018 IEEE International Conference on Healthcare Informatics (ICHI)*. 2018:443-444. doi:10.1109/ICHI.2018.00092

27. Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2016:207-212. Accessed June 26, 2025. https://aclanthology.org/P16-2034.pdf

28. Kramer JH, Mungas D, Possin KL, et al. NIH EXAMINER: Conceptualization and Development of an Executive Function Battery. *J Int Neuropsychol Soc*. 2014;20(1):11-19. doi:10.1017/S1355617713001094

29. Ree A, Bendas J, Pabel L, Croy I, Sailer U. Right between the eyes: Corrugator muscle activity tracks the changing pleasantness of repeated slow stroking touch. *Physiology & Behavior*. 2020;222:112903. doi:10.1016/j.physbeh.2020.112903

30. Welhaf MS, Wilks H, Aschenbrenner AJ, et al. Naturalistic assessment of reaction time variability in older adults at risk for Alzheimer's disease. *Journal of the International Neuropsychological Society*. 2024;30(5):428-438. doi:10.1017/S1355617723011475

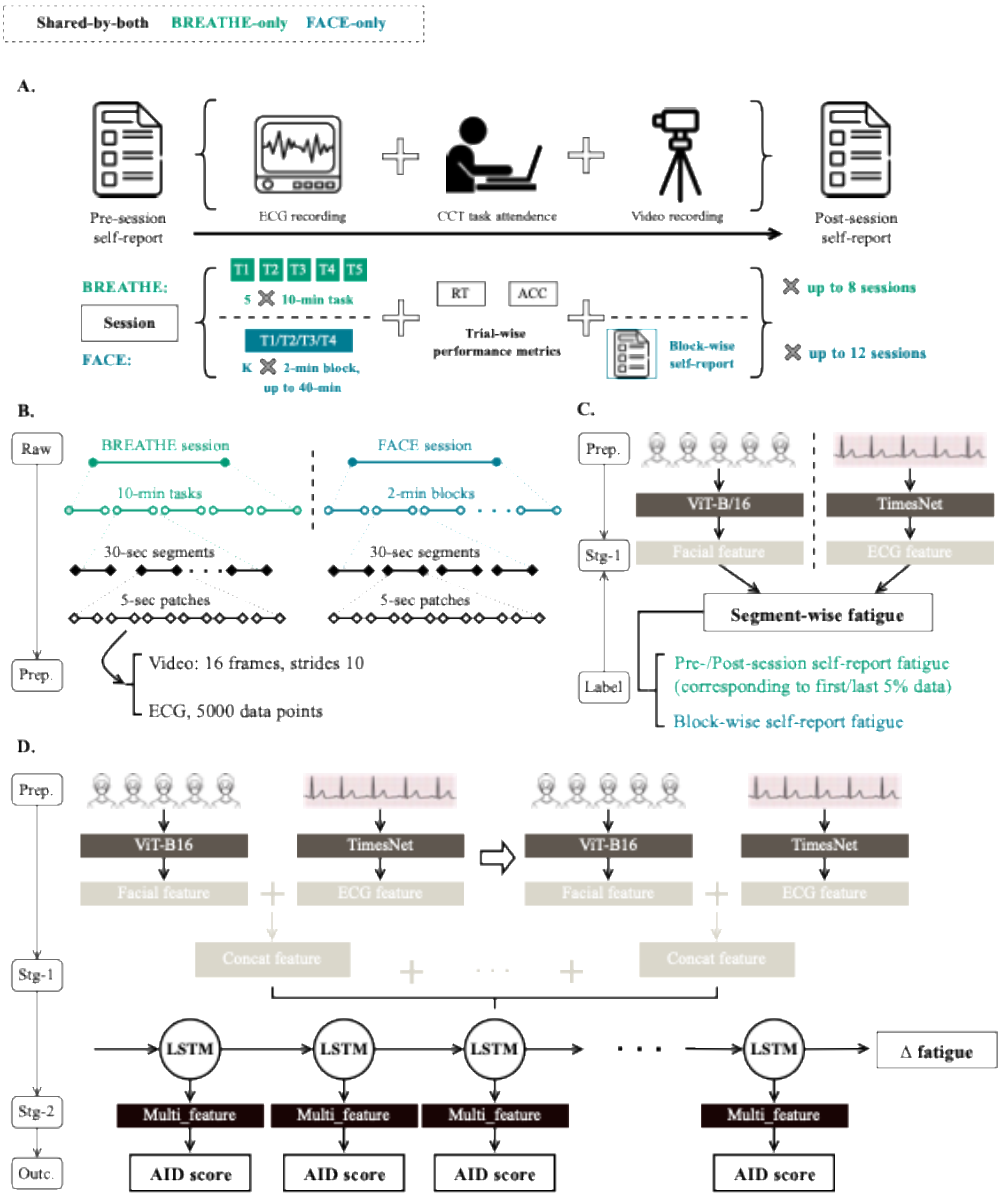

Figure 1. Overview of CCT engagement data collection protocol, AID framework architecture, and processing workflow across BREATHE and FACE datasets. (A) Study protocols in terms of data collection across video, ECG, and CCT tasks; T1-T5 CCT tasks in BREATHE (green) refer to Multiple object tracking, Rapid serial visual presentation, Visual Search, Visual Sweep, and Useful Field of View – in BREATHE, a mix of tasks was used within a session; T1-T4 CCT tasks in FACE (teal) refer to Motion of Target, Sound Sweeps, Mixed Signals, and Delayed Task Switching – in FACE, a single task was used within a session; see details in **Supplemental Materials**. RT= reaction time; ACC= accuracy of the tasks; (B) Processing workflow for extracting facial video and ECG segments; (C) Training stage of AID's video and ECG encoders, respectively; (D) Overall AID framework, integrating video and ECG encoders with a Long Short-Term Memory (LSTM) based fusion module that were used in predicting Δfatigue and generating AID score. What was not shown in the figure was that the top 10% of AID scores within each session were used to select representative 30-second multimodal features (Multi_feature) for post-hoc analysis. Across (B)-(D), Raw=raw data, Prep.=preprocessing stage, Stg-1=modality-specific encoders, Stg-2=multimodal fusion module, Label=labels used for encoder training, Outc.=prediction stage with multimodal fusion and temporal modeling, Concat=concatenation.

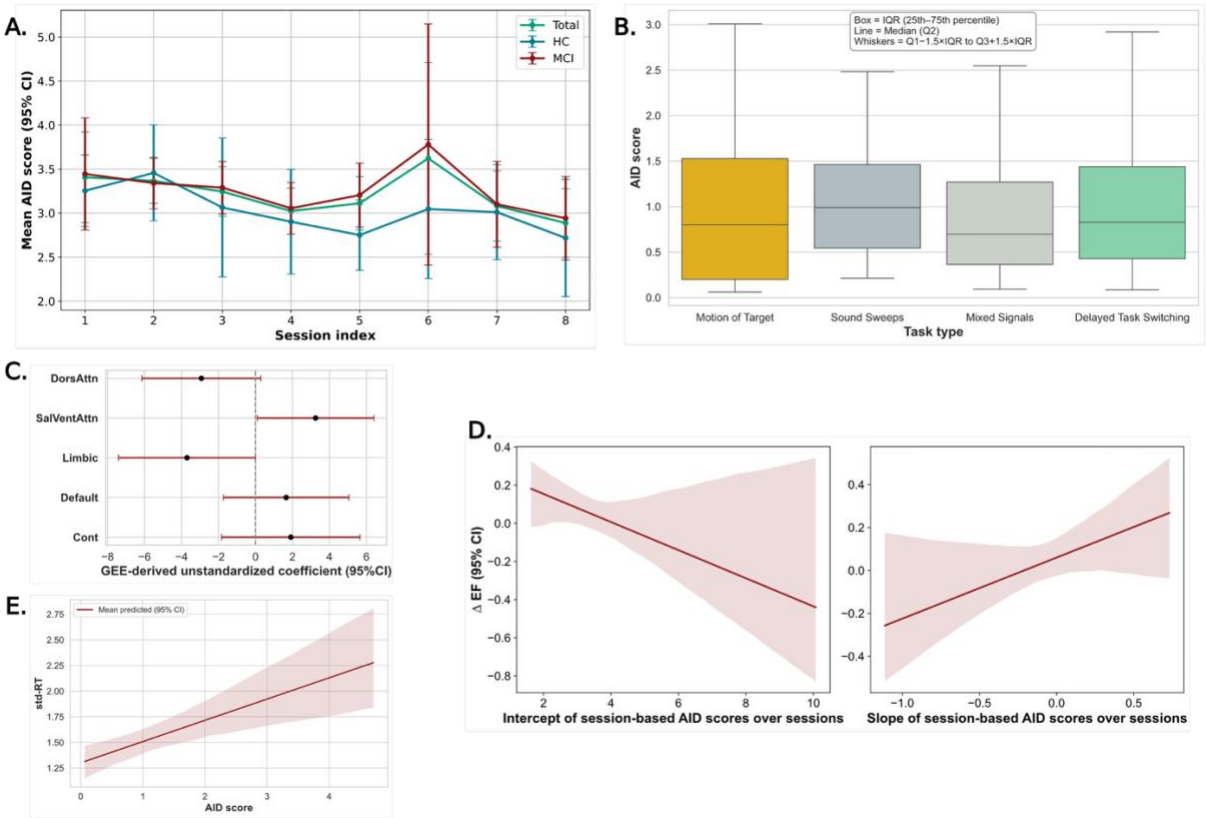

Figure 2. Results of AID framework evaluation and clinical validation. (A) No significant difference in AID score over sessions as a total sample or by cognitive status (HC=healthy control; MCI=mild cognitive impairment); (B) No significant differences in AID scores across task types; (C) Baseline brain network strengths (SalVentAttn and Limbic) predict AID scores in uncorrected tests, which did not survive multi-comparison correction; DorsAttn=Dorsal attention network, SalVentAttn=Salience/Ventral network, Limbic=Limbic network, Default= Default mode network, Cont=Cognitive control network; (D) Smaller intercept of AID scores and greater slopes across sessions predict greater improvements in executive function (ΔEF) when controlling for intervention group assignment; (E) AID scores was significantly related to std-RT, adjusting for session and task type.

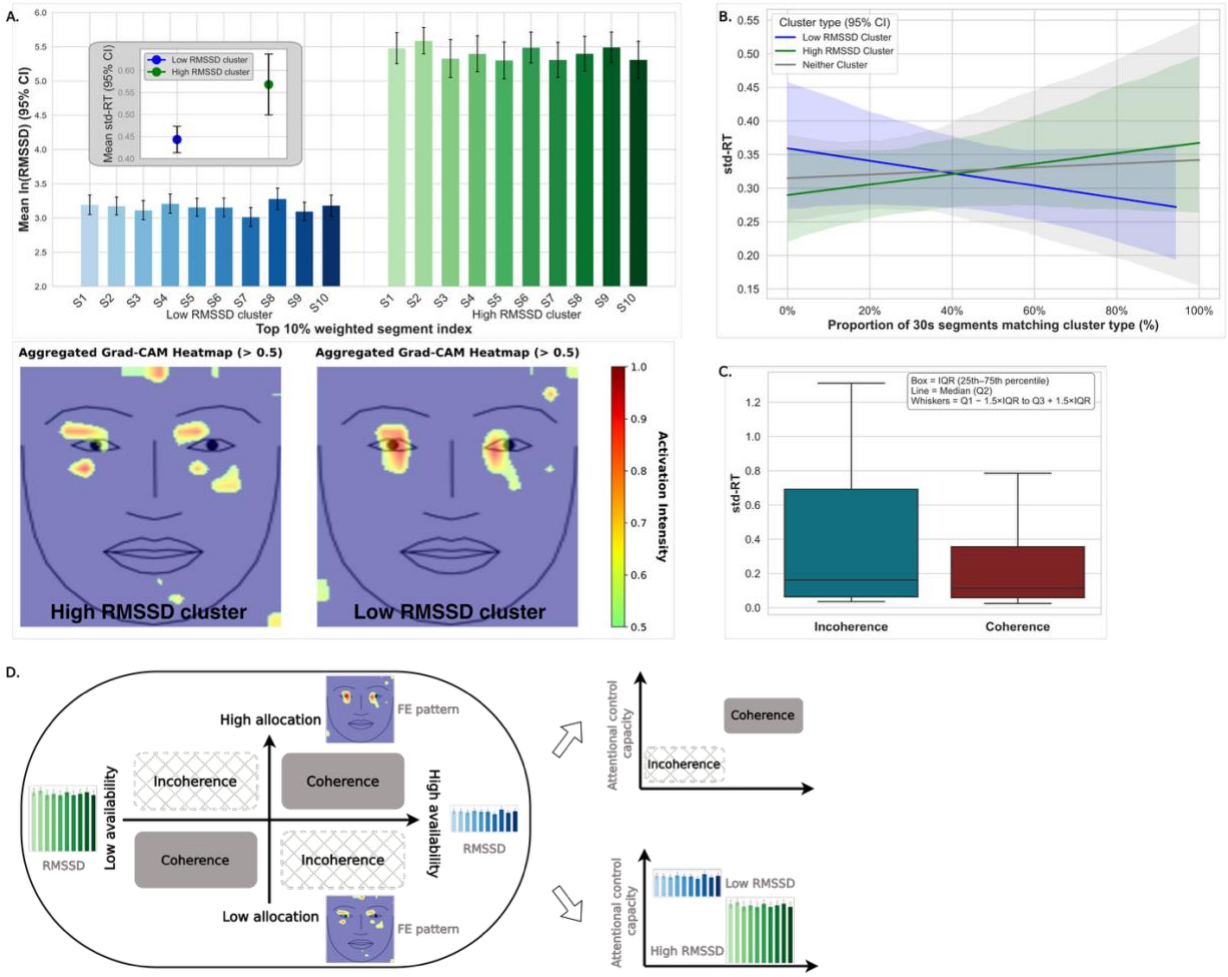

Figure 3. Results of AID framework post hoc interpretation. (A) Two distinct coherent engagement clusters identified from AID's multimodal features differed in RMSSD, facial activation, and attentional control outcome (std-RT). (B) Proportions of segments matching specific RMSSD clusters relate to std-RT, with higher "Low RMSSD" alike proportion associated with better attention (i.e., lower std-RT), while "High RMSSD"-alike proportion trend toward poorer performance; (C) Sessions dominated by coherent RMSSD and FE patterns show significantly better attention (i.e., lower std-RT) than session dominated by incoherent patterns; (D) A two-dimensional framework illustrates attentional control outcome (indexed by session-based std-RT) as a function of availability (ECG-derived RMSSD) and allocation (FE patterns): better attentional control is reflected in both high coherence between availability and allocation, as well as high availability alone.

Table 1. Samples' baseline characteristics overview.

	BREATHE (N=50)	**FACE (N=20)**
Age (years), Mean±SD	74.3±7.12	72.6±7.13
Education (years), Mean±SD	16.2±2.15	17.1±1.92
Sex (Male:Female), %	18:32, 36%	8:12, 40%
Ethnicity (Hispanic/Latino:Not), %	1:49, 2%	0 (1 unknown):19, 0%
Race (White:Others), %	44:5, 90% (1 unknown)	8:12, 40%
MCI:HC, %	40:10, 80%	7:13, 35%
MoCA, Mean±SD	24.5±2.4	24.4±3.0
RAVLT Delayed Recall Raw Score, Mean±SD	5.48±3.0	8.61±3.5
EXAMINER[28], Mean±SD	0.27±0.6	N/A
GDS, Mean±SD	1.98±0.22	6.1±6.25

NOTE: MCI=mild cognitive impairment; HC=healthy control with subjective cognitive complaint; MoCA=The Montreal Cognitive Assessment; RAVLT= Rey Auditory Verbal Learning Test; GDS= Geriatric Depression Scale (NOTE: BREATHE used 15 items and FACE used 30 items).

Table 2. AID performance in predicting Δfatigue across modalities and datasets.

Dataset	Face	ECG	Face+ECG	Significance
BREATHE (Acc)	0.796	0.760	0.817	Face+ECG vs. Face (t=-2.396, p =0.027) Face+ECG vs. ECG (t=-2.572, p =0.019)
BREATHE (F1)	0.792	0.756	0.810	Face+ECG vs. Face (t=-2.721, p =0.014) Face+ECG vs. ECG (t=-2.522, p =0.021)
FACE (Acc)	0.720	0.709	0.732	–
FACE (F1)	0.728	0.693	0.735	–
FACE Validation (Acc)	–	–	0.692	–
FACE Validation (F1)	–	–	0.675	–

NOTE: Accuracy and F1 performance of Face-only, ECG-only, and multimodal Face+ECG models across BREATHE and FACE datasets using Leave-One-Subject-Out cross-validation. Statistical significance (paired t-test) reflects paired comparisons between multimodal and unimodal models. FACE validation results are reported without significance testing, as they reflect held-out generalization performance rather than within-fold comparisons.