AI-assisted modeling of attentional control for intervention engagement

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

Affiliation:

1. Department of Psychiatry and Behavioral Sciences, Stanford School of Medicine, Stanford, CA
2. School of Nursing, University of Rochester Medical Center, Rochester, NY
3. Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY

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

Acknowledgement: Data collection was supported by NIH AG081723, NR015452, and AG084471; and data analysis was supported by Stanford HAI seed funding.

Abstract

Cognitive training is one of the most used cognitive interventions, aiming for preventing or slowing cognitive decline in older adults. However, its effect is inconsistent, often due to the lack of effective engagement. Here we developed a novel AI-assisted multimodal framework, called Attentional Control Index using Digital measures (AID), which combines video-based facial expression analysis with ECG-derived heart rate variability (HRV) to assess the availability and allocation of attentional resources during cognitive training. Using two independent datasets of older adults at elevated risk for cognitive impairment, we demonstrated that AID accurately predicts both behavioral markers of engagement and cognitive outcomes from cognitive training. Additionally, we identified interpretable facial expressions and HRV features that reflect attentional control. These findings establish a robust, generalizable digital marker for assessing, and potentially in the future, enhancing effective engagement in cognitive training by focusing on the real-time dynamics of individual users' attentional control.

1. Introduction

1.1. Limitations of cognitive training and the need to assess engagement

Digitally delivered non-pharmacological interventions represent a promising frontline for preventing or slowing cognitive decline and brain aging in individuals at risk for Alzheimer's disease (AD). 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 and prevent dementia due to its scalability, affordability, and safety profile[1]. However, evidence regarding the effectiveness of CCT remains mixed, in part due to substantial variability in adherence and engagement[2–5]. A key challenge is balancing cognitive demand with the engagement experience: tasks that are too demanding, due to excessive novelty or difficulty, may lead to frustration and dropout, while overly simple tasks with low demand may be ineffective or boring. Both scenarios can fail to promote the desired cognitive plasticity or to produce clinically meaningful transfer effects beyond trained effect in CCT[6].

An emerging priority in the CCT field is understanding *effective engagement*, i.e., a state of sustained energy investment encompassing focus and enjoyment in CCT that drives desirable training outcomes[7,8]. Specifically, there is a push to understand and encourage states that shape cognitive plasticity towards CCT transfer effects[9–12]. Yet current methods for assessing engagement remain limited, especially in older adults at risk for AD. Traditionally, engagement has been measured using adherence (e.g., total time spent on training, number of sessions per week, length of individual sessions, etc.) or task performance within CCT sessions. However, they often fail to reliably predict transfer effects, suggesting they do not capture effective engagement[8]. Moreover, these measures mainly assess the consequences of engagement rather than the dynamics of the engagement process itself, and do not reflect how nonlinear engagement dynamics might shape cognitive plasticity. Self-report fatigue is another commonly used metric, but increasing the frequency of fatigue assessments during a session might more accurately reflect an individual's engagement status while potentially disrupting effect CCT training. Also, the reliability of self-reports decreases as cognitive impairment worsens. These challenges in the literature on CCT and effective engagement underscore a pressing need for more dynamic and real-time assessment tools that can capture the nuances of user engagement as it unfolds.

1.2. Attentional control and its relevance to CCT

Attentional control, the capacity to regulate attention by integrating top-down goals with bottom-up stimuli, is central to cognitive training outcomes[13]. This function is closely tied to brain networks that play separable but complementary roles in regulating attention (e.g., dorsal and ventral attention networks)[13]. In the context of CCT, attentional control 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, as well as longer timescales, while allocation changes moment-to-moment in response to task demands[14]. However, they also interact such that availability can change instantly in response to specific environmental cues, and allocation is also influenced by fluctuating arousal. Thus, two conditions may be particularly relevant for effective engagement: (1) when available resources are fully allocated; (2) when available

resources can be expanded to meet increased task demands. Conventional tools struggle to capture these complex nonlinear interactions, especially in aging populations with elevated AD risk that show increased variability and noise in attentional control[15].

1.3. A new framework using AI-assisted multimodal measures

Recent advances in artificial intelligence (AI) offer a transformative approach to capturing the dynamic and multimodal nature of cognitive engagement during CCT[16]. Traditional measures, such as self-reports or performance-based measures, often fail to capture the subtle, moment-to-moment shifts in attention and effort that influence learning and cognitive plasticity[17,18]. In contrast, AI models, such as deep neural networks and temporal encoders, are well-suited for uncovering complex, nonlinear patterns in high-dimensional data streams, such as facial video or/and psychophysiological recordings, either independently or in combination[19–21].

Prior work, including our own, has shown that the autonomic nervous system (ANS) reflects neural resource availability during cognitively demanding tasks in older adults at risk for AD[22–24]. For example, parasympathetic activity, indexed by ECG-derived Root Mean Square of successive differences (RMSSD) or high-frequency heart rate variability (HRV), reflects the availability of brain resources. Typically, a suppression-and-rebound parasympathetic activity pattern is observed during cognitive training tasks: 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[24,25] (also see our review[22]). 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[26], reflecting the dynamic reallocation of cognitive resources[27]. By fusing facial and psychophysiological signals, AI may assist the development of a fine-grained and interpretable index of engagement that captures not only the consequences but also the insights of process of effective engagement as they evolve throughout CCT tasks.

1.4. Objective and hypotheses

In this work, we proposed a novel framework for quantifying the real-time dynamics of attentional control and enhancing the accuracy of predicting effective engagement in CCT through AI-assisted analysis of biobehavioral data. Specifically, we developed **AID**, an **A**ttentional control **I**ndex using **D**igital measures, leveraging a vision transformer model (ViT-B/16[28]) for facial expression (FE) analysis and a time-series modeling approach (TimesNet[29]) for ECG to quantify real-time attentional control. We hypothesized that a multimodal approach integrating FE and ECG might more accurately reflect effective engagement than either modality alone. Here self-report fatigue throughout CCT sessions was served as a standard measure linking CCT process (i.e., attentional control dynamics during CCT) and consequences (i.e., cognitive improvement from CCT)[30–32] (also see our review[33]), and used to guide the development of AID model.

We aimed to (1) develop and validate AID as a multimodal digital measure-based model for predicting effective engagement [i.e., change of self-report fatigue (Δfatigue) and CCT clinically meaningful outcome: change in executive function(ΔEF)] during or after CCT sessions and examine patterns derived from AID scores across CCT sessions and tasks in datasets collected from two separate studies; and (2) identify interpretable FE and HRV patterns from AID that correspond to neural and cognitive profiles of attentional control [(i.e., brain networks related to

attentional control, and CCT session or task-based standard deviation of reaction time to accurate task responses (std-RT)].

The two datasets used in the study, i.e., BREATHE and FACE, are two intervention studies, using similar attentional control-oriented CCT tasks, sample [mixing older adults being mild cognitive impairment (MCI) and cognitively healthy (HC) with subjective cognitive complaints; see **Table 1** sample characteristics], and CCT engagement related data collection protocols but different study designs (BREATHE – an ongoing three-arm randomized control trial; FACE – a completed single-arm quasi-experiment; see **Figure 1A** and study design details in Methods). Of note, since BREATHE is an ongoing study with intervention assignment being blinded, we only controlled for the intervention group assignment across post-hoc analyses instead of comparing group difference, which served our objective of the present work on developing and evaluating AID as a model for effective engagement instead of evaluating intervention effect on AID.

2. Results

2.1. Develop and validate AID framework by examining multimodal measures of FE and ECG in relation to self-report fatigue in the BREATHE and FACE datasets

Figures 1B-D show data processing and the two-stage analytical flows (i.e., "Modality-Specific Encoders" and "Multimodal Fusion Module") across datasets BREATHE and FACE, with details explained in the **Methods** section. Model evaluations followed the discovery and validation procedure. In the *discovery phase,* we applied leave-one-subject-out (LOSO) cross-validation on the **BREATHE** dataset to evaluate within-cohort generalizability. To further examine performance in a smaller independent sample, we also conducted a separate LOSO analysis within the **FACE** dataset. In both cases, all components of the AID framework, including modality-specific encoders and the multimodal fusion module, were trained and tested strictly within each dataset, with no data overlap across folds. In BREATHE, the multimodal model achieved an accuracy of 0.82 (F1=0.810) for classifying Δfatigue [vs. 0.80 for FE-only (F1=0.79), 0.76 for ECG-only (F1=0.76)]. Similar results were discovered in the FACE dataset [0.73 accuracy for multimodal model (F1=0.74) vs. 0.72 for FE-only (F1=0.73) and 0.71 for ECG-only (F1=0.69)].

In the *validation phase,* we assessed cross-cohort transferability by pretraining the full AID framework on **BREATHE** and fine-tuning it end-to-end on **50% of the FACE** data. Model performance was then evaluated on the **remaining 50% of FACE**. Fine-tuning included both encoder and fusion components, initialized with pretrained weights from **BREATHE**. All training and hyperparameter tuning procedures were conducted independently within each dataset to prevent data leakage. The fine-tuned AID model reached an accuracy of 0.69 (F1=0.68).

Across both BREATHE and FACE datasets, the multimodal model consistently outperformed unimodal models in both accuracy and F1 score for predicting within-session Δfatigue, with multimodal model further validated in FACE, highlighting the complementary roles of FE and ECG measures for engagement.

2.2. Examine patterns derived from AID scores between CCT sessions in BREATHE dataset

Over-session patterns. We first examined the multimodal measure-based AID score derived from the AID framework every 30-second segment and presented as session-based values by averaging all segments-based scores within a session in the BREATHE dataset (see segment

details in **Figure 1B** and AID score generation in Methods). Using Generalized Estimating Equation (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**).

Relationships with baseline brain profile of attentional control. 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*). When examining the relationship between AID scores and baseline brain networks potentially related to attentional control (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 2B**); note that when applying false discovery rate correction, significant results did not survive. When applying cognitive status 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* in our case.

Relationships with CCT outcomes after interventions. We used ΔEF (i.e., the difference between EF after and before intervention) 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 AID scores and ΔEF (session×AID score: Wald's χ^2=7.85, p=0.005). For interpretation purpose, we calculated the intercept and the linear slope of the AID score based on its relationship with the session and plotted the relationships between the intercept or slope of AID score and ΔEF controlling for group. Smaller intercept of over-session AID score (B=-0.07, SE=0.03, t=-2.07, p=.043) and greater slope (B=0.31, SE=0.15, t=2.04, p=.046) were related to greater improvement in EF (see **Figure 2C**).

Visualization of the feature space of the AID framework. To support interpretability and within-session feature analysis, we used corresponding 30-second windows for both FE and ECG analyses, aligning these with the same-window AID scores for evaluation. For downstream clustering (e.g., Figure 2D), we selected the top 10% of 30-second segments within each session based on their aggregated attention scores derived from the LSTM's temporal attention algorithm[34], which quantifies each segment's relative contribution to predicting Δfatigue. These top-ranked segments were assumed to represent periods of peak attentional control and effective engagement. The corresponding multimodal features (concatenated FE and ECG features) from these top-ranked segments were extracted and pooled across sessions and subjects. We first used RMMSD derived from ECG features for k-means clustering due to ANS' role in reflecting attentional control availability. Based on the RMSSD clustering, we also extracted corresponding FE patterns that reflected corresponding allocation. **Figure 2D** displays two coherent engagement profiles based on k-means clustering of the top-ranked 30-second aggregated RMSSD values and corresponding FE patterns (see RMSSD calculation and K-means details in **Methods**). The first cluster was characterized by lower RMSSD and greater facial activation in

the eye and gaze regions, which may reflect focused engagement (labeled as "Low RMSSD" cluster). The second cluster showed higher RMSSD and more activation in the corrugator and cheek regions, possibly indicating disengagement or fatigue (labeled as "High RMSSD" cluster). Importantly, participants in "High RMSSD" cluster showed significantly greater session-based std-RT (i.e., poorer outcome of attentional control) than "Low RMSSD" cluster.

2.3. Examine patterns derived from AID scores between CCT tasks in FACE dataset

Between-task patterns. Following the same procedure described above, we calculated the task-based AID scores for the FACE dataset (see details in **Figure 1B** and **Methods**). 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 3A**).

Relationship with attentional control. Using GEE model taking session-based std-RT as outcome and session, task type, and AID scores as predictors (*Equation 2.2*), there was significant relationship between AID scores and std-RT (B=0.10, SE=0.03, Wald's χ^2=10.93, p<0.001), suggesting smaller AID scores related to better outcome of attentional control (std-RT) (see **Figure 3B**).

Interpreting engagement profiles learned from BREATHE to FACE dataset. We calculated the proportion of 30-second segment RMSSD and FE patterns that were alike or unlike the two clusters ("Low RMSSD" and "High RMSSD" clusters) identified from BREATHE (see **Methods**) and created two types of metrics for each session: (1) proportion scores of RMSSD reflecting how frequently a session's segments resembled the "Low RMSSD", "High RMSSD" clusters, or neither; (2) a binary coherence variable indicating whether a session was dominated by aligned RMSSD and FE patterns (coherence) or not (incoherence), with "coherence" defined as sessions in which more than 30% of segments were concurrently assigned to the same cluster type (either "Low" or "High") across both RMSSD and FE patterns. First, using the GEE model (*Equation 2.3a*) taking std-RT as outcome, and sessions, task type, and proportion score of RMSSD resembling either cluster as predictors, sessions with RMSSD patterns resembling "Low RMSSD" clusters had lower/better std-RT compared to those without such patterns with a trend (B=-0.09, SE=0.05, Wald's χ^2=3.39, p=0.066); sessions with "high RMSSD" cluster-alike patterns had greater/worse std-RT compared to those without such patterns (B=0.11, SE=0.14, Wald's χ^2=3.80, p=0.051) (**Figure 3C**). Next, using the GEE model (*Equation 2.3b*) taking std-RT as outcome, and session, task type, and coherence vs. incoherence variable as predictors, sessions dominated by coherent patterns had significantly lower/better std-RT compared to those dominated by incoherent pattern (B=-0.14, SE=0.06, Wald's χ^2=5.43, p=0.020) (**Figure 3D**). Together, these results support our prior findings in the BREATHE dataset, demonstrating that RMSSD patterns, reflecting availability of attentional control, were critical for differentiating session-based attentional control; and further, incoherent patterns between RMSSD and FE might further worsen attentional control compared to coherent patterns.

3. Discussion

3.1. Overall observations

By integrating video-derived FE patterns and ECG-derived HRV patterns, the AID framework quantifies real-time attentional control states in 30-second intervals. These states are associated with fatigue-related disengagement and preserved attentional control capacity and can be used to predict changes in executive function after intervention. The AID framework is robust, in terms of having non-significant variation, against individual differences in demographics (i.e., age, sex, and education) or cognitive status (i.e., HC vs. MCI). Importantly, we confirmed the attentional control model by focusing on the availability of resources (via HRV), as well as coherent patterns of availability and allocation (via corresponding FE and HRV patterns). We identified two coherent patterns: one with low RMSSD and focused eye gaze, and another with high RMSSD and activity in the corrugator and cheek regions, corresponding to high or low attentional control, respectively. Overall, AID provides a robust and interpretable framework for capturing effective engagement during computerized cognitive training.

3.2. Role of AID score in evaluating effective engagement

The literature highlights the challenge of linking performance in CCT sessions to improvements in transfer effects—cognitive outcomes not directly practiced during CCT sessions (see our review[21]). Our BREATHE data corroborates this finding, revealing no relationship between outcome of attentional control practiced in CCT sessions (measured by session-based std-RT) and changes in EF post-intervention. In contrast, session-based AID scores, which are related to both behavioral and neural aspects of attentional control, were also predictive of changes in EF. Specifically, across two datasets, lower AID scores indicated poorer attentional control capacity. Furthermore, in the BREATHE study, a lower initial AID score (intercept of over-session AID score) coupled with significant improvements in AID score (slope) were associated with greater EF enhancements, suggesting more pronounced benefits for those starting with weaker attentional control.

It is important to note that in the current BREATHE sample, there was no significant change in AID score across sessions when controlling for intervention group assignment. However, due to the blinding of the intervention groups, 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 the intervention blinding in the BREATHE study is lifted, and as future CCT interventions incorporate real-time monitoring of FE and ECG using the AID framework, we anticipate improved prediction of transfer effects both at the individual and session levels.

3.3. Interpretation of AID-informed FE and ECG patterns for quantifying properties of attentional control

We confirmed our earlier perspective that effective engagement is a multimodal construct[8]. The multimodal model outperformed unimodal models in predicting changes in self-report fatigue across both datasets. We also need to discuss separate scenarios related to the proposed availability (HRV) and allocation (FE) relationship in the context of attentional control. In our case, availability-allocation coherence (as opposed to incoherence) and high availability of resources (as opposed to low availability) are equally important when evaluating individual- or momentary-level attentional control. See **Figure 4** for illustration.

In terms of specific patterns of HRV and FE, we found that high attentional control capacity was associated with low RMSSD and FE patterns focusing on eye and eye gaze. Notably, in the literature on healthy younger adults, RMSSD, an index of the parasympathetic nervous system (PNS), typically shows a positive relationship with cognitive capacity. However, the

contradictory finding here aligns with emerging literature on older adults with cognitive impairment and our proposed framework and earlier work[22,35]. 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. We found consistent results across the BREATHE and FACE datasets. Further, in terms of FE patterns -- eye and eye gaze vs. corrugator and cheek regions in the literature reflect distinct behavioral meanings in terms of cognitive vs. affective regulation[35,36].

3.4. Limitations and next steps

While promising, our findings should be interpreted in light of several limitations. First, the datasets represent a relatively modest and demographically narrow sample. Larger and more diverse cohorts are needed to validate generalizability. Second, although AID scores were predictive of EF change, their relationship to other cognitive domains (e.g., memory) remains to be explored, so as AID's causal influence on cognitive improvement. As mentioned above, when BREATHE blinding is lifted, along with future studies may test real-time adaptive training paradigms informed by AID feedback. Finally, based on our post-hoc feature analyses, given the role of eye gaze, future work may explore the temporal granularity of AID scoring, integrate additional signals (e.g., pupillometry).

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 CCT engagement by tailoring it to attentional control dynamics among older adults with AD risk.

4. Methods

4.1. Protocols and participants

We used two independent datasets, i.e., BREATHE and FACE, each comprising older adults at risk for AD who completed attentional control tasks over multiple CCT sessions. Both studies included synchronized facial video and ECG data collection to support multimodal modeling of engagement.

BREATHE (Breathing, Relaxation, Attention Training & Health in Older Adults) protocol is an ongoing randomized control trial comparing CCT-only, CCT following paced breathing, and imagery-based control among older adults with MCI and HC. The intervention comprises weekly in-lab and daily at-home sessions across 8 weeks. During 50-minute in-lab sessions, participants, regardless of intervention assignment, completed 5 computer tasks targeting attentional control and processing speed, with facial video and ECG continuously recorded throughout each session. Of note, BREATHE is an ongoing clinical trial without unblinding its group assignment. When evaluating over-session or after intervention effect, we only controlled for group assignment without specifying any between-group difference.

FACE (Facial Expression-based Personalization Engine) protocol is a CCT in-lab quasi-experiment designed to evaluate attentional control and fatigue in older adults with AD risk. Participants completed up to six 40-minute sessions, each featuring a single attentional control task while facial video and ECG were recorded continuously. Within each session, 2-minute blocks of tasks were provided. The attentional control tasks were conceptually overlapping with BREATHE (targeting the same cognitive processes) with different user interface designs.

Across both datasets, during each CCT session, participants underwent synchronized collection of facial video and ECG signals to capture biobehavioral data of attentional control-oriented engagement. Facial videos were recorded using a front-facing webcam, with frames sampled at 16 frames per 5-second clip and resized to 160×160 pixels. Facial regions were detected and cropped using RetinaFace[37] for downstream analysis. ECG signals were recorded via a 3-lead setup using BioLab v3 (MindWare Technologies Ltd.) at 1000 Hz. Raw ECG traces were converted to a single-column format and processed using NeuroKit2[38] for artifact removal. This integrated pipeline ensured temporal alignment and quality control across both modalities for robust multimodal modeling.

Details of participant eligibility related to MCI diagnosis across two studies are detailed in BREATHE protocol[39]. Of note, HC proportion of BREATHE protocol was added later using same MCI group related data collection protocol.

Figure 1A shows the CCT protocol setup across BREATHE and FACE. A description of the total 9 attentional control tasks was provided in **Supplemental Materials**.

4.2. Measures

Self-report fatigue: Across both datasets, a 6-item visual analogue scale was administered at the beginning and end of each session to assess perceived fatigue, and averaged with higher values indicating more fatigue. In FACE, there was also an item of "tiredness" for every 2-minute block within sessions. The fatigue data was used in two ways: (1) For training and evaluating modality-specific encoders (*Stage 1*), binary fatigue labels were generated by converting the session-level mean score (ranging from 0 to 5) into a binary format using a cutoff of 2 [fatigue (≥2) vs. no fatigue (<2)]. Note for BREATHE, the first and last 5% of each session were weakly labeled using the pre/post-session scores, while the middle 90% remained unlabeled and contributed to unsupervised temporal representation learning. For FACE, each 2-minute task block was treated as an independent unit, with the block-level "tiredness" rating serving as the ground truth. Segment-level fatigue labels were then assigned based on the block to which they belonged, allowing finer-grained encoder supervision. These labels supervised encoder learning specific to either FE or ECG modality (see **Figure 1C** and **Method 4.3**). (2) For training and evaluating the multimodal fusion module developed in *Stage 2*, change in fatigue (Δ fatigue = post-session - pre-session score) was used as the binary classification target, representing shifts in attentional control across sessions (see **Figure 1D and Methods 4.3**).

Attentional control-related measures: Across two datasets, behavioral outcome of attentional control was quantified as the standard deviation of RT (std-RT) over trial-based RT to accurate responses during each CCT session. Higher value indicated worse attention. In BREATHE, resting-state fMRI at baseline was used to derive functional connectivity strength within five canonical attention-related networks (i.e., dorsal attention, salience/ventral attention, limbic, default mode, and control), following Yeo's 7-network parcellation (see **Supplement Materials**).

Transfer effect from CCT: EF was assessed at baseline and post-intervention across all participants in BREATHE. A composite score was calculated based on the NIH EXAMINER established algorithm[40], and a change score was calculated as the difference between post-intervention and baseline.

4.3. AID Framework development

The AID framework was developed to predict session-based change of fatigue (Δ fatigue: post – pre), serving as a proxy for effective engagement. 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-1D** illustrate the AID framework's architecture and data partitioning strategy.

*Stage 1: Modality-Specific Encoders (see **'Stg-1' in Figure 1C and 1D**)*

FE encoder: We employed a Vision Transformer (ViT-B/16)[28], pretrained on the FERV39k Multi-Scene Dataset[41], 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[29], a temporal foundation model optimized for time-series data, pretrained on ECG Heartbeat Categorization Dataset[42–44]. 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 (facial expression and ECG) was trained to predict fatigue levels using a binary classification task. Binary self-reported fatigue was used. 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 attentional control, 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[34,45], 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 fatigue change (defined as post-session minus pre-session fatigue score). Sessions with an increase of ≥1 point in fatigue were labeled as "fatigue increase," and others as "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.

The model was implemented in PyTorch 2.1.2 with an NVIDIA A100 GPU. Training followed a cross-subject protocol, holding out one subject for testing while using the remaining data for training. The model was optimized using Adam with a learning rate of 0.0005 and cross-entropy loss for classifying session-level fatigue changes with early stopping based on validation loss. A batch size of 32 was used, with each batch consisting of 30-second multimodal sequences constructed from aligned FE and ECG features. Training ran for up to 20 epochs, with structured logging and output saving for reproducibility. The final model used two LSTM layers with a hidden size of 128, a dropout rate of 0.3, and attention-weighted pooling for temporal window aggregation. These settings were selected based on the best average F1 score across validation folds. Pretrained encoder weights were kept frozen during initial fusion training and then unfrozen in a second phase for full fine-tuning. All experiments were seeded to ensure consistent results across runs.

Score calculation. To enable interpretable and segment-wise analysis of fatigue, we computed the AID scores per 30-second block. These scores were used for post-hoc analysis of fatigue for AID's interpretation and hypothesis verification. **AID score**: Mean of LSTM outputs from concatenated FE + ECG features across the 30-second block. This reflects the integrated temporal signature of fatigue-related engagement.

4.4. Other data analyses related to validating and interpreting AID Framework

Generalized estimating equation (GEE) models with AR(1) correlation matrix were used for all post-hoc modeling, including session- or task-level variability in AID scores; relationships between AID scores and attentional control outcome (std-RT), baseline brain network strength, and EF outcomes as a function of session or task-level AID dynamics.

Equations included:

Equations for BREATHE data

1.1. $Y^{session}_{Multi_score}$=Intercept+$X_{session}$+$X_{group}$+$\varepsilon$;

1.2. $Y^{session}_{Multi_score}$=Intercept+$X_{session}$+$X_{cognitive_status}$+$X_{group}$+$\varepsilon$;

1.3. $Y^{session}_{Multi_score}$=Intercept+$X_{session}$+$X_{dorsal_attention}$ +$X_{ventral_attention}$+X_{limbic}+X_{defaul_mode} +$X_{control}$+ε; (NOTE: extreme head motion related outliers, n=2, were excluded from the sample)

1.4. $Y^{session}_{Multi_score}$=Intercept+$X_{session}$+$X_{age}$ +X_{sex}+$X_{education}$+ε;

1.5a. $Y_{\Delta EF}$=Intercept+$X_{session}$+$X^{session}_{std_RT}$ +$X_{session} \times X^{session}_{std_RT}$ +X_{group}+ε;

1.5b. $Y_{\Delta EF}$=Intercept+$X_{session}$+$X^{session}_{std_RT}$ +$X_{session} \times X^{session}_{std_RT}$ +$X^{session}_{Multi_score}$ +$X_{session} \times X^{session}_{Multi_score}$+$X_{group}$+$\varepsilon$;

Equations for FACE data

2.1. $Y^{session}_{multi_score}$=Intercept+$X_{session}$+$X^{session}_{task}$+$\varepsilon$;

2.2. $Y^{session}_{std_RT}$=Intercept+$X_{session}$+$X^{session}_{task}$ +$X^{session}_{Multi_score}$+$\varepsilon$;

2.3a. $Y_{std_RT}^{session}$=Intercept+$X_{session}$+$X_{task}^{session}$ +$X_{proportion_RMSSD}^{session}$+$\varepsilon$; (NOTE: "proportion_RMSSD" refer to session-based RMSSD patterns similar to "low RMSSD" cluster or "high RMSSD" cluster in BREATHE data)

2.3b. $Y_{std_RT}^{session}$=Intercept+$X_{session}$+$X_{task}^{session}$ +$X_{coherence_status}^{session}$+$\varepsilon$

RMMSD calculation and k-mean clustering RMSSD was computed using NeuroKit2[38]'s *'hrv_time()'* function, applied to 30-second ECG segments after R-peak detection and preprocessing. Specifically, we first applied *'ecg_peaks()'* with artifact correction to identify R-peaks, then computed RMSSD by passing the detected peaks to *'hrv_time(peaks, sampling_rate=1000)'* and extracting the RMSSD value for each segment.

For post-hoc RMSSD pattern analysis in the BREATHE dataset, we performed k-means clustering using the aggregated RMSSD values. The optimal number of clusters was determined using within-cluster sum of squared errors (elbow method) and silhouette scores, both of which indicated that two clusters provided the best solution (see **Supplementary Figures** S1–S2 for cluster selection).

References

1. 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

2. Lampit A, Hallock H, Valenzuela M. Computerized Cognitive Training in Cognitively Healthy Older Adults: A Systematic Review and Meta-Analysis of Effect Modifiers. Gandy S, ed. *PLoS Med*. 2014;11(11):e1001756. doi:10.1371/journal.pmed.1001756

3. Hoorelbeke K, Vervaeke J, Siegle GJ, Baeken C, Koster EH. Individual differences associated with treatment adherence and transfer effects following gamified web-based cognitive control training for repetitive negative thinking. *Internet interventions*. 2022;27:100507.

4. Evers A, Klusmann V, Schwarzer R, Heuser I. Improving cognition by adherence to physical or mental exercise: A moderated mediation analysis. *Aging & mental health*. 2011;15(4):446-455.

5. He Z, Tian S, Singh A, et al. A Machine-Learning Based Approach for Predicting Older Adults' Adherence to Technology-Based Cognitive Training. *Information Processing & Management*. 2022;59(5):103034. doi:10.1016/j.ipm.2022.103034

6. Turnbull A, Seitz A, Tadin D, Lin FV. Unifying framework for cognitive training interventions in brain aging. *Ageing Research Reviews*. 2022;81:101724. doi:10.1016/j.arr.2022.101724

7. Nahum-Shani I, Yoon C. Towards the Science of Engagement with Digital Interventions. *Curr Dir Psychol Sci*. 2024;33(4):239-246. doi:10.1177/09637214241254328

8. 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

9. Csikszentmihalyi M. FLOW: The Psychology of Optimal Experience. Published online 2000.

10. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. *Behav Med Pract Policy Res*. 2017;7(2):254-267. doi:10.1007/s13142-016-0453-1

11. Kelders SM, Van Zyl LE, Ludden GD. The concept and components of engagement in different domains applied to eHealth: a systematic scoping review. *Frontiers in psychology*. 2020;11:926.

12. Short CE, Rebar AL, Plotnikoff RC, Vandelanotte C. Designing engaging online behaviour change interventions: a proposed model of user engagement. *Eur Health Psychol*. 2015;17(1):32-38.

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

14. 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

15. 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

16. 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

17. 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

18. Marcotte TD, Schmitter-Edgecombe M, Grant I. *Neuropsychology of Everyday Functioning*. Guilford Publications; 2022. Accessed June 30, 2025. https://books.google.com/books?hl=en&lr=&id=gcBZEAAAQBAJ&oi=fnd&pg=PA1998&dq=self-reports+or+performance-based+measures,+often+fail+to+capture+the+subtle,+moment-to-moment+shifts+in+attention+and+effort+that+influence+learning+and+cognitive+plasticity&ots=qO_aAr4wIU&sig=PzKcQn3aU9yTfudProjHdpxeC5s

19. Zhao G, Li Y, Xu Q. From Emotion AI to Cognitive AI. *IJNDI*. Published online December 22, 2022:65-72. doi:10.53941/ijndi0101006

20. Li Y, Liu Y, Nguyen A, et al. Interactions for Socially Shared Regulation in Collaborative Learning: An Interdisciplinary Multimodal Dataset. *ACM Trans Interact Intell Syst*. 2024;14(3).

21. Tutuianu GI, Liu Y, Alamäki A, Kauttonen J. Benchmarking deep Facial Expression Recognition: An extensive protocol with balanced dataset in the wild. *Engineering Applications of Artificial Intelligence*. 2024;136:108983. doi:10.1016/j.engappai.2024.108983

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

23. Peralta-Malváez L, Turnbull A, Anthony M, Adeli E, Lin FV. CCA identifies a neurophysiological marker of adaptation capacity that is reliably linked to internal locus of control of cognition in amnestic MCI. *GeroScience*. 2023;45(3):1803-1815. doi:10.1007/s11357-023-00730-8

24. 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

25. 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

26. 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

27. Babaei E, Srivastava N, Newn J, Zhou Q, Dingler T, Velloso E. Faces of Focus: A Study on the Facial Cues of Attentional States. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Association for Computing Machinery; 2020:1-13. doi:10.1145/3313831.3376566

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

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

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

31. Ren P, Anderson AJ, McDermott K, Baran TM, Lin F. Cognitive fatigue and cortical-striatal network in old age. *Aging*. 2019;11(8):2312-2326. doi:10.18632/aging.101915

32. Lin F, Ren P, Cotton K, Porsteinsson A, Mapstone M, Heffner KL. Mental Fatigability and Heart Rate Variability in Mild Cognitive Impairment. *The American Journal of Geriatric Psychiatry*. 2016;24(5):374-378. doi:10.1016/j.jagp.2015.12.012

33. 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

34. 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

35. 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

36. 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

37. Deng J, Guo J, Ververas E, Kotsia I, Zafeiriou S. Retinaface: Single-shot multi-level face localisation in the wild. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. ; 2020:5203-5212. Accessed July 1, 2025. http://openaccess.thecvf.com/content_CVPR_2020/html/Deng_RetinaFace_Single-Shot_Multi-Level_Face_Localisation_in_the_Wild_CVPR_2020_paper.html

38. Makowski D, Pham T, Lau ZJ, et al. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behav Res*. 2021;53(4):1689-1696. doi:10.3758/s13428-020-01516-y

39. 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

40. 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

41. 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

42. Bousseljot R, Kreiseler D, Schnabel A. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. 1995;40(s1):317-318. doi:10.1515/bmte.1995.40.s1.317

43. Moody GB, Mark RG. The impact of the MIT-BIH Arrhythmia Database. *IEEE Engineering in Medicine and Biology Magazine*. 2001;20(3):45-50. doi:10.1109/51.932724

44. 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

45. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural computation*. 1997;9(8):1735-1780.

46. Nasreddine ZS, Phillips NA, Bédirian V, et al. The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment. *Journal of the American Geriatrics Society*. 2005;53(4):695-699. doi:10.1111/j.1532-5415.2005.53221.x

47. Schmidt M. *Rey Auditory Verbal Learning Test*. Western Psychological Services Los Angeles; 1996.

48. Montorio I, Izal M. The Geriatric Depression Scale: A Review of Its Development and Utility. *International Psychogeriatrics*. 1996;8(1):103-112. doi:10.1017/S1041610296002505

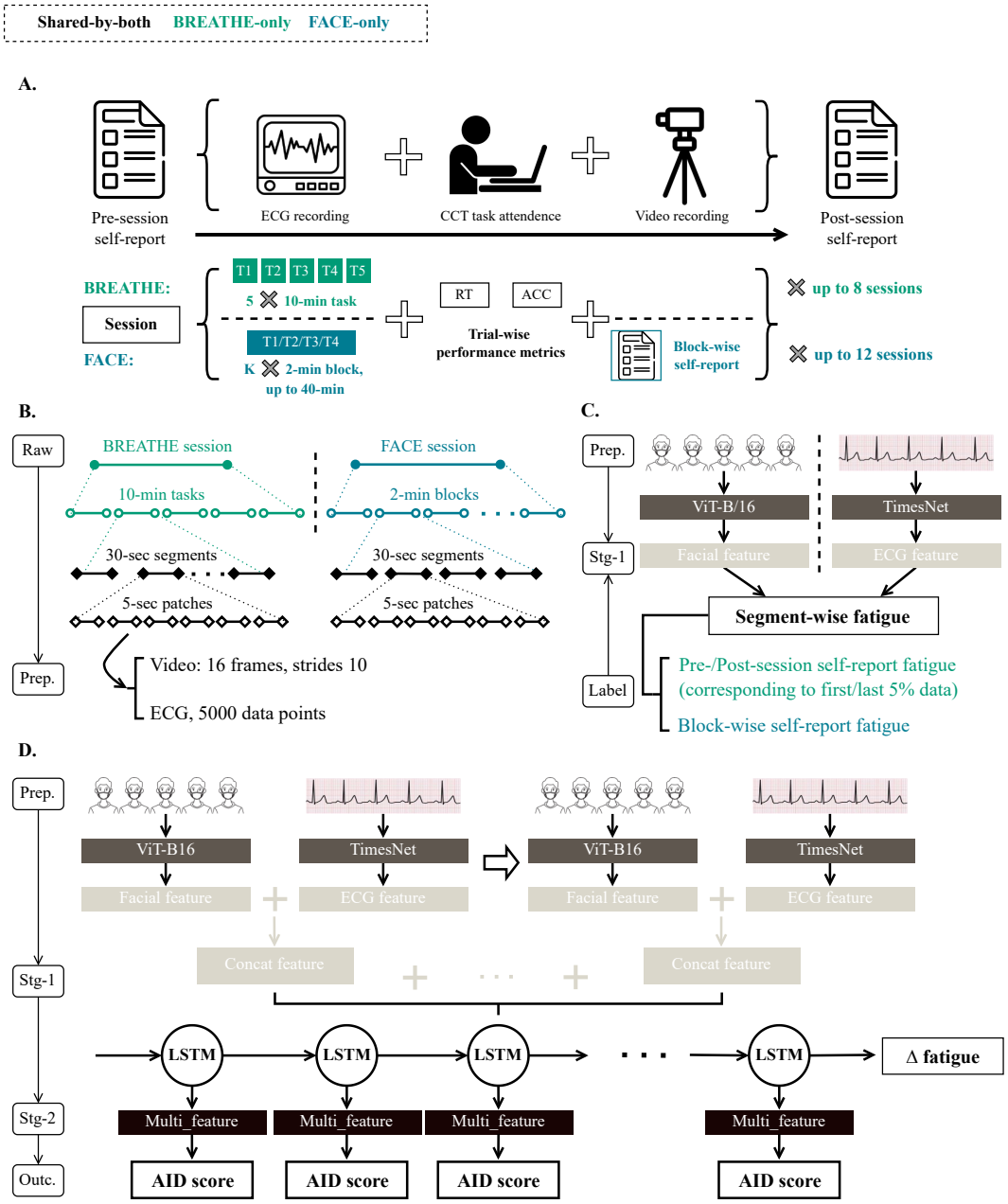


Figure 1. Overview of CCT engagement data collection protocol, AID framework architecture, and processing workflow across BREATHE and FACE datasets. (A) Study protocols in terms of data collection across video, ECG, and CCT tasks; T1-T5 CCT tasks in BREATHE (green) refer to Multiple object tracking, 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.

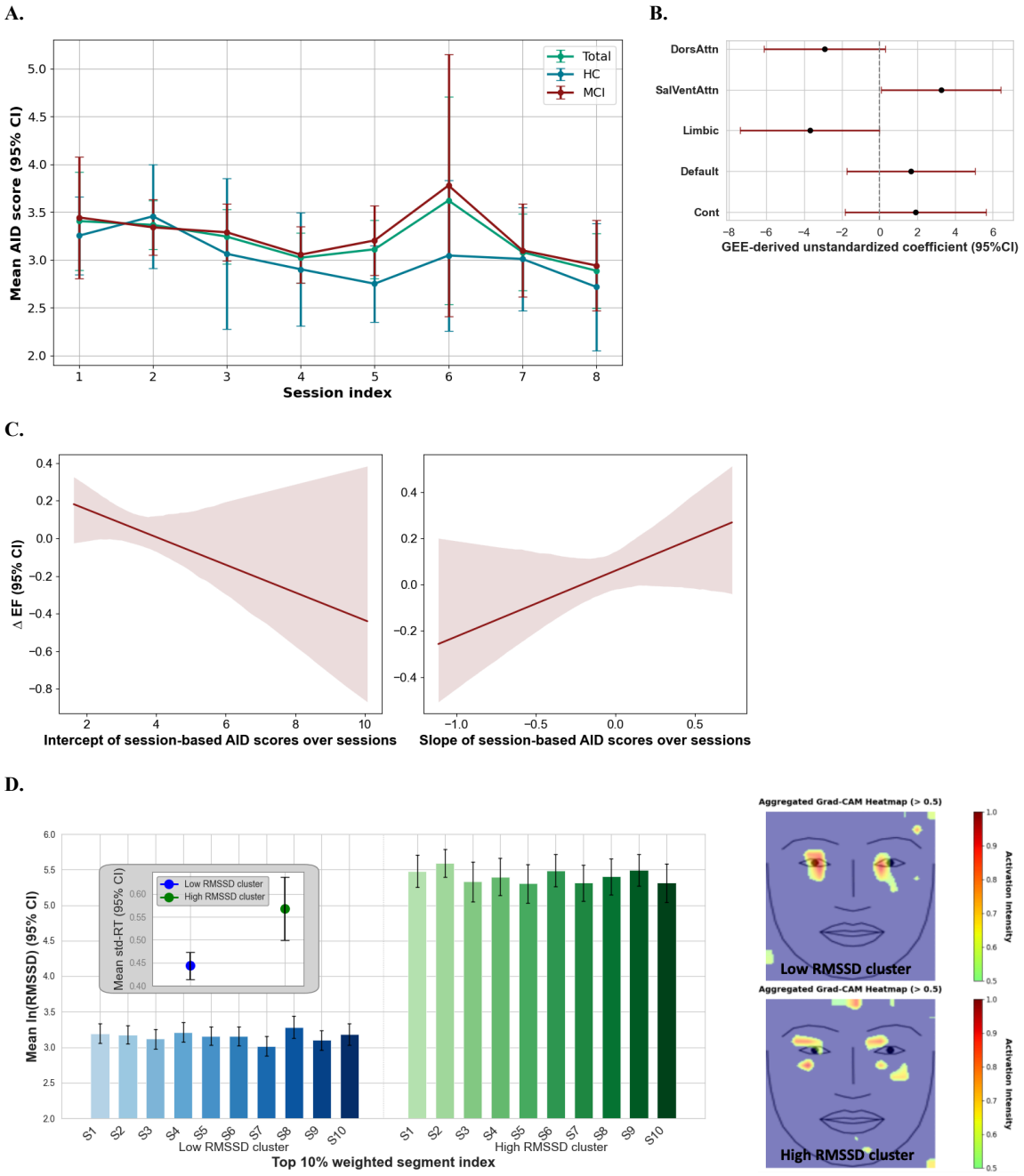


Figure 2. Discovery of session-level AID performance in BREATHE. (A) No significant difference in AID score over sessions as a total sample or by cognitive status (HC=healthy control; MCI=mild cognitive impairment); (B) Baseline brain network strengths (SalVentAttn and Limbic) predict AID scores in uncorrected 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; (C) Smaller intercept of AID scores and greater slopes across sessions predict greater improvements in executive function (ΔEF) when controlling for intervention group assignment; (D) Two distinct coherent engagement clusters identified from AID's multimodal features differed in RMSSD, facial activation, and attentional control outcome (std-RT).

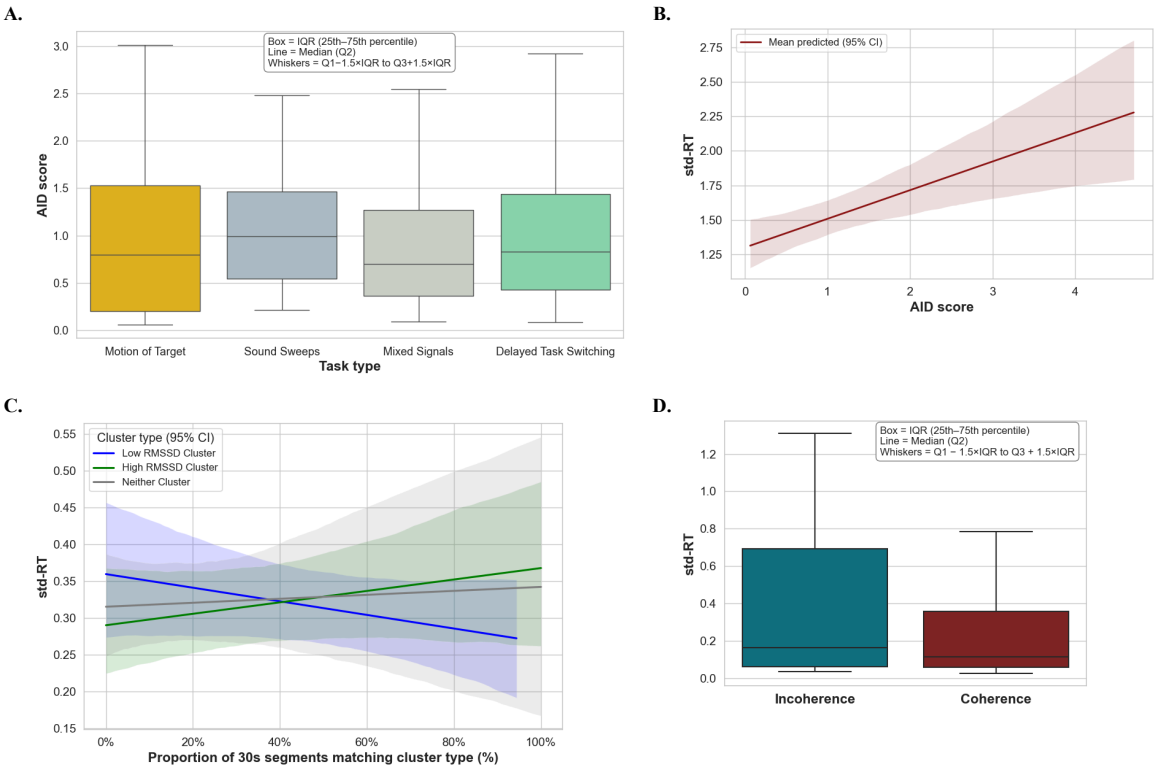


Figure 3. Validation of task-level AID performance in FACE. (A) No significant differences in AID scores across task types; (B) AID scores was significantly related to std-RT, adjusting for session and task type; (C) Proportions of segments matching specific RMSSD clusters relate to std-RT, with higher "Low RMSSD" alike proportion associated with better attention (i.e., lower std-RT), while "High RMSSD"-alike proportion trend toward poorer performance; (D) Sessions dominated by coherent RMSSD and FE patterns show significantly better attention (i.e., lower std-RT) than session dominated by incoherent patterns.

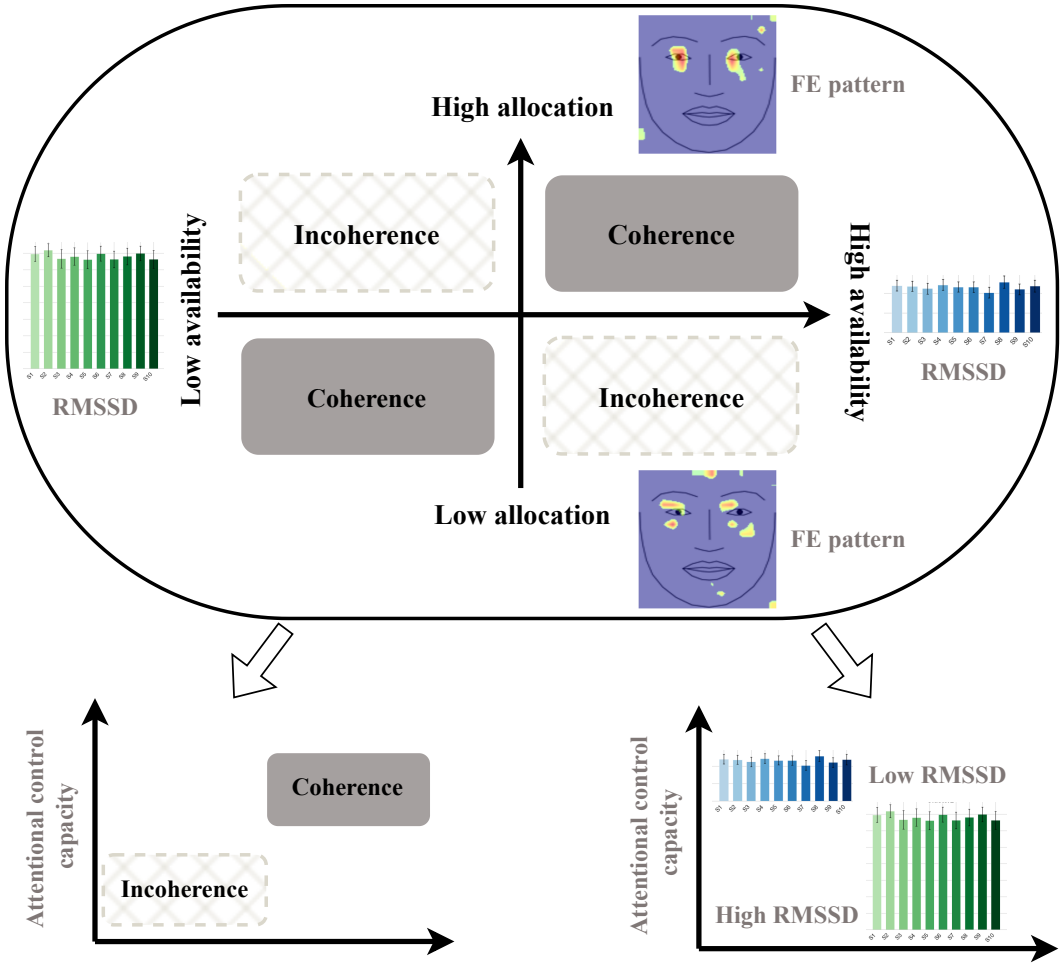


Figure 4. Illustration of a multi-dimensional attentional control framework. A two-dimensional framework illustrates attentional control outcome (indexed by session-based std-RT) as a function of availability (ECG-derived RMSSD) and allocation (FE patterns); better attentional control is reflected in both high coherence between availability and allocation, as well as high availability alone.

Table 1. Samples' baseline characteristics overview.

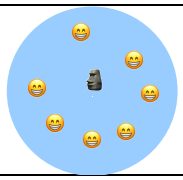
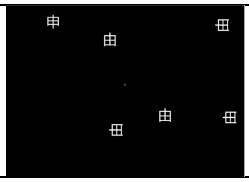
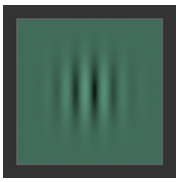
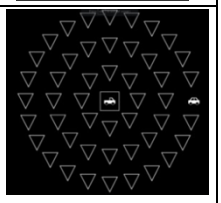
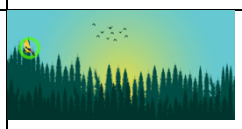
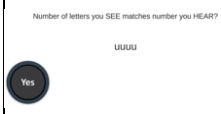
	BREATHE (N=50)	**FACE (N=20)**
Age (years), Mean±SD	74.3±7.12	72.6±7.13
Education (years), Mean±SD	16.2±2.15	17.1±1.92
Sex (Male:Female), %	18:32, 36%	8:12, 40%
Ethnicity (Hispanic/Latino:Not), %	1:49, 2%	0 (1 unknown):19, 0%
Race (White:Others), %	44:5, 90% (1 unknown)	8:12, 40%
MCI:HC, %	40:10, 80%	7:13, 35%
MoCA[46], Mean±SD	24.5±2.4	24.4±3.0
RAVLT[47] Delayed Recall Raw Score, Mean±SD	5.48±3.0	8.61±3.5
EXAMINER[40], Mean±SD	0.27±0.6	N/A
GDS[48], 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).

Supplementary materials list:

1. Description of nine attentional control tasks;
2. fMRI network calculation procedure;
3. Cluster number determination via Elbow and Silhouette analyses (Figures S1–S2).

Supplemental materials

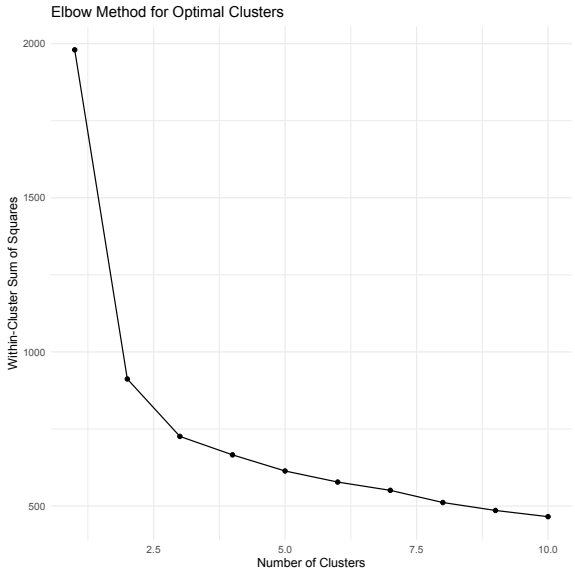
1. Description of CCT tasks	
tasks used in BREATHE	
Multiple object tracking (MOT) aims at improving visuospatial working memory capacity. Participants track a varying number of target objects (defined by their spatiotemporal locations) among visually identical distractors.	
Rapid serial visual presentation (RSVP) assesses visual processing speed. Participants view a stream of images in rapid succession and must determine whether a target image appears as quickly as possible.	
Visual search targets visual attentional processing. Participants must locate a target stimulus among a field of other distractor stimuli as quickly as possible.	
Visual sweep targets visual processing speed. Participants observe a spatial frequency sweep (movement of bars) and determine whether the sweep moves inward or outward.	
Useful Field of View (UFOV) targets attentional processing and measures the ability to extract information in a single glance without moving the head or eyes. Participants are tested on central visual processing speed, divided attention between central and peripheral stimuli, and selective attention between peripheral stimuli among distractors.	
tasks used in FACE	
Motion of Target (MOT) aims at improving visuo-spatial working memory capacity. In this multiple object tracking exercise, participants track a varying number of target objects (defined by their spatio-temporal location) among visually-identical distractors. The difficulty levels vary in eccentricity, speed, contrast, and number of objects to track.	
Sound Sweeps targets auditory processing speed. In this time-order-judgement exercise, two successive frequency modulated tone sweeps are presented and participants indicate whether the pitch (i.e., frequency) increased or decreased within each tone. The difficulty levels differ as a function of central frequency, interstimulus interval/sweep ratio, and sweep speed in milliseconds.	
Mixed signal targets multi-sensory attention and cognitive control using the Stroop task training concept. In this exercise, the participant is asked to focus on certain information while ignoring competing information. The exercise requires to listen to a letter, color, number or other piece of information while looking at a set of letters, symbols, numbers, words or other information. The difficulty level is determined by the complexity, similarity of the visuo-auditory pairing, and similarity of competing information.	

Delayed task switching employs a basic task switching design in which participants are presented with a target stimulus (indicated with an arrow) and two options. They must find the best match for the target from the two options based on a specific criterion: for example, color or shape. This task engages task switching/flexibility because the criteria vary from trial to trial, e.g., one trial may ask to match based on color the next based on shape. This task also engages working memory because the criterion for matching is presented a few seconds prior to the trial and is not present on screen during the trial, meaning that participants must store this information in working memory to perform the task.	

2. MRI data acquisition and processing

MRI data were acquired with Siemens 3T Prisma (VE11C), equipped with a 64-channel head coil. *T1w:* A five-minute scan was collected with a single-shot echo-planar imaging (EPI) MPRAGE sequence [repetition time/echo time (TR/TE) = 1400/2.34 ms, slice thickness = 1 mm, resolution = 1 mm isotropic, 192 slices, PE acceleration = GRAPPA, acceleration factor = 2, flip angle = 8°, orientation = sagittal, echo spacing = 7 ms, field of view (FOV) = 256 mm]. *Resting-state fMRI:* A five-minute run was collected with a gradient EPI multiband (MB) sequence (TR/TE = 1010/44 ms, slice thickness = 2 mm, resolution = 2 mm isotropic, R = 1, flip angle = 70°, orientation = transversal, echo spacing = 0.56 ms, FOV = 256 mm, acquisition matrix = 128 x 128, 80 slices, 300 volumes, acceleration factor = 8). Slice acquisition order was interleaved. Participants were instructed to keep their eyes open. Field maps with opposite phase encoding directions were acquired for distortion correction.
The preprocessing of both structural and functional images was performed using fMRIprep pipeline (Esteban et al., 2019), an open-source tool based on Nipype v1.8.6 (Gorgolewski, et al., 2011). The T1 structural images underwent preprocessing steps including skull-stripping, segmentation, surface construction, and volume-based spatial normalization. Preprocessing steps of the resting-state functional images include head motion correction, slice timing correction, susceptibility distortion correction (SDC), distortion co-registration to T1 images, and spatial normalization to standard space. SDC was performed via estimated fieldmaps from two echo-planar imaging references with TOPUP (Andersson et al., 2003). For the preprocessed functional images, additional denoising steps were performed with Nilearn v0.11.0 (https://github.com/nilearn/nilearn): despiking was applied first to remove the artifact spikes in the BOLD signal with 3dDespike from AFNI (https://afni.nimh.nih.gov/afni/). Twelve head motion parameters and 5 noise components from white matter and CSF via aCompCor approach, estimated from fMRIprep, were regressed out as confounding covariates, and then a bandpass filter (0.01, 0.1) was implemented. Lastly, the functional images were smoothed using a 6-millimeter full-width at half maximum Gaussian smoothing kernel. The functional connectivity matrixes were constructed via Schaefer atlast (Schaefer 2018), with 400 regions of interest. The functional connectivity for seven yeo networks was calculated by averaging across the connectivity strengths across all edges within each network.

3. K-mean cluster solution



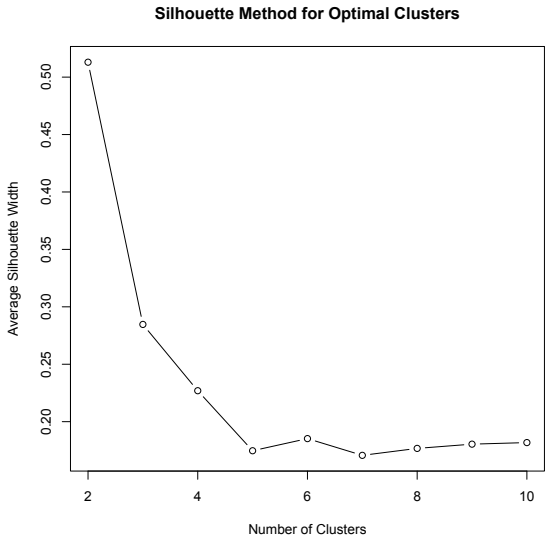


Figure S1 (left): Elbow method analysis based on within-cluster sum of squared errors (WCSS) across different values of k (number of clusters). The plot shows a clear inflection point at k = 2, indicating the optimal number of clusters that balance model complexity with interpretability for multimodal AID features.

Figure S2 (right): Average silhouette scores for clustering performance across different values of k. The highest silhouette score was observed at k = 2, supporting a two-cluster solution that maximizes inter-cluster separation and intra-cluster cohesion of AID features.