CASES: A Cognition-Aware Smart Eyewear System for Understanding How People Read

XIANGYAO QI*, QI LU*, and WENTAO PAN*, School of Computer Science, Fudan University, China
YINGYING ZHAO, School of Computer Science, Fudan University, China
RUI ZHU, Bayes Business School, City, University of London, United Kingdom
MINGZHI DONG and YUHU CHANG†, School of Computer Science, Fudan University, China
QIN LV, Department of Computer Science, University of Colorado Boulder, United States
ROBERT P. DICK, Department of Electrical Engineering and Computer Science, University of Michigan, United States
FAN YANG, School of Microelectronics, Fudan University, China
TUN LU, NING GU, and LI SHANG†, School of Computer Science, Fudan University, China

The process of reading has attracted decades of scientific research. Work in this field primarily focuses on using eye gaze patterns to reveal cognitive processes while reading. However, eye gaze patterns suffer from limited resolution, jitter noise, and cognitive biases, resulting in limited accuracy in tracking cognitive reading states. Moreover, using sequential eye gaze data alone neglects the linguistic structure of text, undermining attempts to provide semantic explanations for cognitive states during reading. Motivated by the impact of the semantic context of text on the human cognitive reading process, this work uses both the semantic context of text and visual attention during reading to more accurately predict the temporal sequence of cognitive states. To this end, we present a Cognition-Aware Smart Eyewear System (CASES), which fuses semantic context and visual attention patterns during reading. The two feature modalities are time-aligned and fed to a temporal convolutional network based multi-task classification deep model to automatically estimate and further semantically explain the reading state timeseries. CASES is implemented in eyewear and its use does not interrupt the reading process, thus reducing subjective bias. Furthermore, the real-time association between visual and semantic information enables the interactions between visual attention and semantic context to be better interpreted and explained. Ablation studies with 25 subjects demonstrate that CASES improves multi-label reading state estimation accuracy by 20.90% for sentence compared to eye tracking alone. Using CASES, we develop an interactive reading assistance system. Three and a half months of deployment with 13 in-field studies enables several observations relevant to the study of reading. In particular, observed how individual visual history interacts

*Equal contribution
†Corresponding authors

Authors' addresses: Xiangyao Qi, 21210240299@m.fudan.edu.cn; Qi Lu, 21212010025@m.fudan.edu.cn; Wentao Pan, 21110240007@m.fudan.edu.cn, School of Computer Science, Fudan University, Shanghai, China, 200438; Yingying Zhao, yingyingzhao@fudan.edu.cn, School of Computer Science, Fudan University, Shanghai, China, 200438; Rui Zhu, rui.zhu@city.ac.uk, Bayes Business School, City, University of London, London, United Kingdom, EC1Y 8TZ; Mingzhi Dong, mingzhidong@gmail.com; Yuhu Chang, yhchang@fudan.edu.cn, School of Computer Science, Fudan University, Shanghai, China, 200438; Qin Lv, qin.lv@colorado.edu, Department of Computer Science, University of Colorado Boulder, Boulder, Colorado, United States, 80309; Robert P. Dick, dickrp@umich.edu, Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan, United States, 48109; Fan Yang, yangfan@fudan.edu.cn, School of Microelectronics, Fudan University, Shanghai, China, 201203; Tun Lu, lutun@fudan.edu.cn; Ning Gu, ninggu@fudan.edu.cn; Li Shang, lishang@fudan.edu.cn, School of Computer Science, Fudan University, Shanghai, China, 200438.

2474-9567/2023/9-ART115 $15.00
https://doi.org/10.1145/3610910

with the semantic context at different text granularities. Furthermore, CASES enables just-in-time intervention when readers encounter processing difficulties, thus promoting self-awareness of the cognitive process involved in reading and helping to develop more effective reading habits.

CCS Concepts: • **Human-centered computing** → **Mobile devices**.

Additional Key Words and Phrases: Smart eyewear, reading, cognition-aware, eye-tracking, visual attention

ACM Reference Format:
Xiangyao Qi, Qi Lu, Wentao Pan, Yingying Zhao, Rui Zhu, Mingzhi Dong, Yuhu Chang, Qin Lv, Robert P. Dick, Fan Yang, Tun Lu, Ning Gu, and Li Shang. 2023. CASES: A Cognition-Aware Smart Eyewear System for Understanding How People Read. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 3, Article 115 (September 2023), 31 pages. https://doi.org/10.1145/3610910

1 INTRODUCTION

Reading is a fundamental approach to learning, through which people can expand their vocabulary, gain knowledge, and develop skills. Research has shown a positive relationship between reading and learning; for example, the more people read, the more effectively they improve vocabulary, knowledge levels, and cognitive skills [15]. In fact, reading has long been considered the most important path to lifelong learning, and lifelong readers are generally more successful, both personally and professionally [24, 76].

The science of reading has attracted decades of interest in human-computer interaction (HCI) [27, 81], cognitive science [40, 44], psychology [67], educational psychology [11, 77], cognition and neuroscience [82], pedagogy [37], and brain science [2, 51]. Reading is a cognitive process and understanding it benefits numerous research communities. Studying how people understand the semantics and syntax of text can aid in understanding natural language representation and processing, which are key functionalities of human-level intelligence [38]. Understanding the reading process can also advance the theory of human behavior, thus benefiting the domains of applied psychology, pedagogy, and educational psychology. For instance, we can scrutinize human cognitive abilities [35] such as verbal working memory capacity, inhibitory control ability, perceptual speed, and immediate and delayed effects on reading processes. Furthermore, understanding how people read sheds light on reading patterns and strategies, potentially helping readers achieve metacognitive awareness and read more efficiently [21, 58, 84]. In particular, HCI researchers have studied enhancing human reading efficiency [80], reading proficiency [52], reading skills [55], reading comprehension performance [34, 43], and reading outcomes [28].

Reading is a multi-level interactive eye-mind cognitive process. In the short term, readers visually perceive each word, encode it, and mentally assign semantics. In the long term, readers visually perceive a sentence and mentally associate it with context and domain knowledge [39]. Reading can be viewed as a sequence of numerous time-varying states. For instance, some studies explored the state of mind wandering, to detect whether a reader is cognitively engaged or decoupled from the current reading task [19, 54]. Furthermore, some researchers studied the state of having difficulty processing unfamiliar words [33, 72]. However, we note that processing difficulties can present at multiple granularities, e.g., readers may encounter difficulties at the level of a single word, a sentence, or a paragraph. Since it is hard to enumerate all reading states, we focus on the problem of probing the reading cognitive process to detect and explain multiple states at word and sentence levels. Specifically, we investigate whether a reader's mind is wandering, whether the reader is positively engaged, and when comprehension is delayed due to word- or sentence-level processing difficulties.

Eye movements are good indicators to infer the cognitive process [1, 64, 74, 83]. This is based on the eye-mind hypothesis [39], which states that there is a close relationship between where the eyes look and where the mind is engaged. Owing to the fast development of eye-tracking technologies, we can easily access eye-tracking data [3, 50] to explore eye-mind relationships. Numerous researchers have extended the relationship between eye movements and cognitive processes [65, 67]. Also, numerous prevalent methods design eye-tracking reading systems to automatically track the participants' eye movements in a non-intrusive way [16, 37, 72, 73].

These works have summarized some hand-engineered eye movement features to probe the reading cognitive process [72, 73].

However, eye-tracking technologies suffer from a number of shortcomings. The error of commercially available eye-tracking technologies typically ranges from 1 to 4 degrees [46, 60]. Under reading scenarios, this angular accuracy translates to a spatial tracking resolution of about 1.4–2.6 cm. Considering a computerized-reading task where the distance from eye to screen is 40–50 cm, this means that the resolution of the eye tracker is about 3 to 4 lines for a single-spaced document and about 1 to 3 words in the horizontal direction. Such low spatial resolution makes it infeasible to track reading states during word-by-word and line-by-line reading because we cannot accurately locate the words and lines. Previous studies tackled this problem by using an unrealistic setting with a very wide line spacing (e.g., triple-spaced [16]), leaving them unsuitable for use with normally spaced text. In addition, eye-tracking techniques are subject to the inherent transient jitter [9] of human gaze and vertical drift, which require constant calibration [7]. Eye-tracking techniques suited to real-world scenarios have the potential to advance the study of reading.

Furthermore, existing methods ignore contextual influences from text, resulting in less accurate reading state estimation and undermining semantic explanation for these states. Given the same reading context and motivations, the factors influencing reading states mainly pertain to the reading material's and subject's domain knowledge about the content. For example, a good reader may cross-reference previously read text to assist in understanding new and unfamiliar text [33]. In such cases, the high reading frequencies of the earlier text do not necessarily imply that they are difficult. To correctly estimate the current reading state, it is important to be aware of the semantic meaning of the current text, the cross-referenced text, their semantic correlations, and real-time eye gaze patterns. However, it is a non-trivial task to properly fuse the semantics of reading text and eye movements and learn from them in progressive reading scenarios, and it is more challenging to infer semantic explanations for reading state timeseries.

This work aims to provide accurate estimations and semantic explanations for reading state timeseries to support research and outreach efforts in the field of reading science. To this end, we pose the following two research questions (RQ) and posit the corresponding hypotheses.

RQ1: Do readers in the same reading states show different visual attention distributions on the reading text?

Hypothesis 1: Readers in the same reading state will show varying visual attention histories (detailed in Section 3), e.g., different total fixation duration, reading times, number of fixations, etc. That is, the visual attention histories of readers in the same reading state differ from each other.

RQ2: When readers are in the same reading states, e.g., encountering difficulty progressing, how does reader visual attention interact with semantic cues in the text?

Hypothesis 2: As indicated by previous studies [19, 73], readers' cognitive effort in processing text is positively related to the difficulty of the text. However, in contrast with previous studies, we further hypothesize that readers can overcome reading difficulties by fetching contextual semantic cues from the surrounding text. When progress is blocked, easy text that is semantically related to difficult text also receives more visual attention and cognitive effort.

The motivation for this work is that the semantic context of text has a direct impact on the multi-level interactive eye-brain cognitive reading process. Leveraging the rich semantic information about reading materials, which can be extracted by advanced natural language processing (NLP) techniques [61, 87], can improve estimation accuracy and provide semantic interpretation of reading states. The semantic information is high-resolution because NLP models can provide semantics at the word level [61, 87]. The inherent hierarchical structure of the semantic information can also be inferred by summarizing the semantics of words to a sentence level. The high-resolution semantic information can compensate for the low-resolution eye movements for more accurate reading state estimation. More importantly, the real-time interaction of eye movements and semantic context can provide semantic explanations for the ongoing reading states.

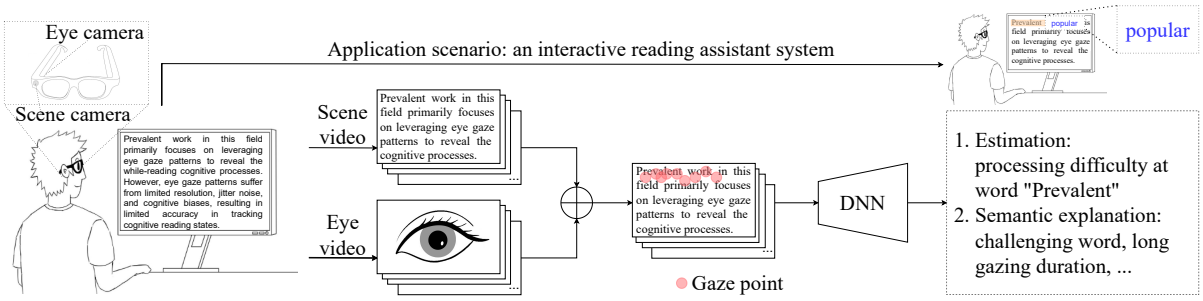


Fig. 1. The proposed CASES smart eyewear system.

To this end, we present a Cognition-Aware Smart Eyewear System (CASES) capable of measuring reading (cognitive) state timeseries. Figure 1 illustrates the overview of the proposed system. At the heart of CASES is a bi-modal multi-task network named CASES-Net, which takes the bi-modal data, i.e., the eye-tracking and reading text data, as inputs and estimates cognitive reading states in real-time at two granularities: word and sentence level. To collect high-quality bi-modal data, CASES uses two cameras to record the two required modalities automatically: an outward-facing scene camera to capture text and an inward-facing camera to track gaze points during reading. CASES is implemented in the form of eyewear to avoid interfering with the reading process when collecting data. Surveys are deferred until after a reading task is completed, also to avoid interference.

CASES-Net employs a four-layer temporal convolutional network (TCN) based module to fuse the two types of sequential modalities, one of which is informed by semantic information extracted from the pre-trained NLP models [6, 86]. We treat estimations at two granularities as two distinct but related tasks and propose a shared convolutional filter mechanism within the TCN to learn the characteristics of the two tasks and their commonalities. Moreover, we design a multi-task and hierarchical loss function to guide reading state estimation. To evaluate CASES, we first collect and construct a dataset and then demonstrate that CASES has higher reading state estimation accuracy than baseline methods. To sum up, CASES-Net combines gaze and semantic information to better estimate reading states. More importantly, it provides semantic explanations for these reading states. The well-trained deep model can automatically detect when users encounter reading difficulty without requiring further inputs (e.g., feedback) from them, thereby limiting potential subjective biases.

This work makes the following contributions.

- We present a Cognition-Aware Smart Eyewear System (CASES) to probe and explain human cognitive processes while reading. CASES aims to support the study of reading and learning to read, as well as supporting HCI and educational applications investigations on improving reading productivity. The CASES system is equipped with a deep neural network, CASES-Net, that extracts features pertaining to the visual attention history and text semantic content. It fuses the two types of features via a shared convolutional filter mechanism based on TCN to enable accurate reading state estimation at various granularities.
- CASES is evaluated in real-world contexts. We conduct an ablation study involving 25 participants, in which CASES delivered superior reading state detection to baseline methods. Specifically, encoding text semantic content facilitates learning from context cues and improves reading state estimation accuracy. Compared with the conventional eye-tracking-only method, we improve accuracy by 20.90% for sentence. Furthermore, the text semantic context enables quantitative explanations of reading (cognitive) states.
- We integrate CASES into a novel interactive reading assistant system. Three and a half months of deployment with 13 in-field studies demonstrate that the integrated system can enable helpful interventions for readers,

thus improving self-awareness in the reading process and helping readers adopt more effective reading habits.

The rest of this paper is organized as follows. Section 2 surveys related work. Section 3 clarifies the key concepts used in this work. Section 4 details the proposed network and our built real-time reading state detection and intervention system. Section 5 presents the experimental setups and results. Section 6 presents our findings when using CASES in practice, general discussion, and future direction. Finally, Section 7 concludes this work.

2 RELATED WORK

This work is mostly relevant to three broad areas: reading science, eye-tracking in reading, and natural language processing.

2.1 Science of Reading

Reading science has attracted decades of interest in various research communities, e.g., HCI, pedagogy, and educational psychology. These studies primarily deal with the outcomes of reading [28] and reading comprehension [43]. Recently, researchers have studied reading patterns and strategies that improve the efficiency of reading [21, 27, 58, 84], e.g., interactive reading systems that detect mind wandering during reading [19, 54]. They mitigated the negative effect of mind wandering on reading comprehension using just-in-time interventions [19, 54]. Other methods detect words readers do not know automatically [27] and provide appropriate help [33, 72]. In psychology, applied psychology, and educational psychology, researchers primarily focused on studying how texts are read and comprehended [11, 62, 67, 77, 83]. For example, Perfetti et al. delivered a blueprint of reading, consisting of the visual process, representation process that converts visual perception into a linguistic representation, and operation process on the representation [62]. In cognition science, neuroscience, and brain science, extensive reading studies focus on developing computational theories of cognition [47]. One important branch studies the representations and processing of natural languages by the human brain [38]. For example, Lewis et al. contributed a theoretical framework to explain how verbal working memory supports sentence processing [47]. Kamide et al. studied how the global and local information in texts impact sentence processing [40]. Cognitive scientists usually jointly consider language representation and processing [23] based on the belief that discovering language representation can help answer questions about computation, and vice versa. Schrimpf et al. provided computationally explicit evidence that language comprehension mechanisms in human brains are fundamentally shaped by predictive processing through an integrative modeling approach [69].

In summary, previous works on the science of reading primarily focus on leveraging eye-tracking during reading to study the reading process and outcomes. However, they focus less on how individual readers perceive and process the text in real time. This study introduces context information from texts to the study of reading cognitive processes.

2.2 Eye-Tracking in Reading

Eye-tracking technology can acquire real-time eye movements in a non-intrusive manner [8]. It is natural to utilize eye movement data to probe the reading process, as the reading process initiates visual input and operates as an interactive eye-mind cognition process [39]. Over the past decades, numerous studies have focused on analyzing eye movement data obtained during reading to understand the reading cognitive process and provide reading assistance [4, 19, 25, 32, 54, 73]. For example, Hyrskykari proposed a gaze-aware reading assistance system to provide help at the right time without interrupting the reader's thoughts [32]. Cheng et al. proposed a social reading system, in which they demonstrated that sharing eye gaze annotations generated by experts promoted reading comprehension for non-experts [10]. Bottos and Balasingam presented an approach to accurately track the horizontal eye-gaze points in reading scenarios [4]. In addition, there are also many studies focused on

detecting reading behaviors, such as mind wandering [19, 54] or encountering difficulties in comprehending unfamiliar words [33, 72].

In general, these relevant methods have demonstrated that eye movement data helps understand the reading cognitive process. However, the semantic information of the text, which is closely related to the reading process, is rarely used in previous studies. This study jointly considers text semantic information and eye movement data can facilitate understanding the reading process and how readers comprehend texts.

2.3 Nature Language Processing

Natural language processing (NLP) uses computational techniques to represent and analyze human languages [12] (see [42] for a comprehensive review). NLP can usually be classified into two categories: natural language understanding and natural language generation. As discussed above, this work uses natural language understanding techniques to obtain semantic contextual information from texts. Successful natural language understanding techniques can provide generic models for NLP downstream tasks, such as analyzing the association among text components [18], extracting keywords [6, 71], and analyzing syntax [48]. For example, Linzen et al. pointed out that, given targeted syntax supervision, a long short-term memory (LSTM) network can learn syntax information [49]. Later, they further stated that linguists and neural network researchers might contribute to each other's areas [48]. Furthermore, NLP neural networks can provide good representations of text; for example, the bidirectional encoder representations from transformers (BERT) model [18], which is based on transformers [79], can obtain state-of-the-art results on several NLP tasks by providing high-quality language representations. Considering the dependency between the masked positions and the discrepancy from pretrain-finetune that BERT neglects, Yang et al. proposed a generalized autoregressive pretraining method to overcome the limitations of BERT [86]. Their pre-trained model, XLNet, outperforms BERT on various tasks. Our work builds on recent progress in NLP by using pre-trained NLP models to help understand the reading cognitive process.

3 PROBLEM FORMULATION

This section clarifies three important concepts used in this work: eye movements, visual attention, and semantic attention.

Eye Movements: Eye movement patterns can reveal reading strategies and are vital to understanding the reading cognitive process. As shown in existing studies [16, 53], reading generally consists of a series of pauses and rapid shifts in gaze locations. The pauses are called fixation, and the shifts are called saccades. These patterns reflect the low-level oculomotor characteristics during reading, typically determined by the physical properties of text, such as the positions or lengths of words.

By exploring eye movement patterns, researchers establish connections between low-level eye movement behaviors and higher-level cognitive processes during reading [74]. First, research shows that the direction and duration of eye fixation reveal how the cognitive process unfolds over time [72, 73]. More specifically, fixation locations indicate the attended content, while fixation duration suggests the level of cognitive effort invested by the reader, i.e., longer fixation suggests more effort. Second, the processing time-course of eye movement patterns is widely used to reveal the temporally continuous reading process, often linked with comprehending or memorizing. For example, one common temporal reading activity is to move the gaze backward to review the already-read content. In this case, the informative eye movement patterns might be the reading and regression durations, also called the second pass [31]. Finally, to alleviate the potential inter-person variations, recent work also designs global features or statistical features based on eye movement patterns to access the reading process, such as the number of saccades, saccade frequencies, and variations in fixation duration [16]. Given the potential ability of eye movement patterns to reveal reading cognitive processes, this work also employs these hand-engineered features as valuable indicators. However, to better suit our case, we first distinguish the

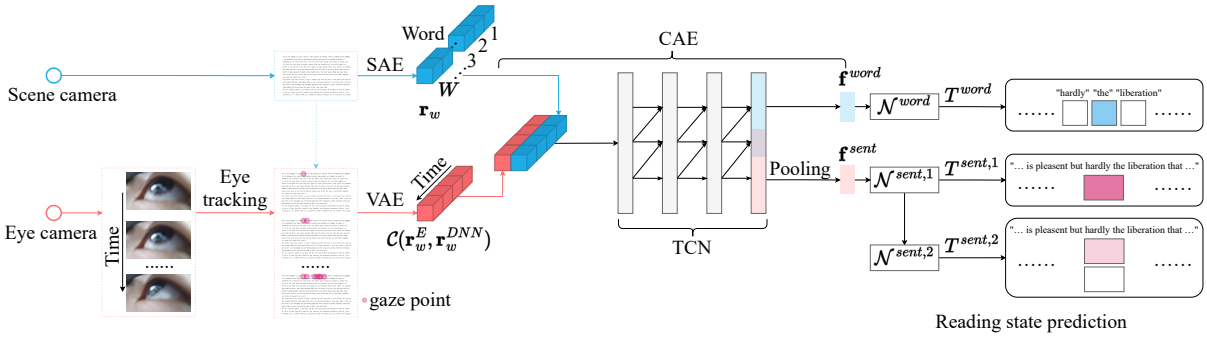


Fig. 2. Overall pipeline of the CASES-Net.

representing eye movement patterns at two granularities and then re-design them at word and sentence levels. More details can be found in Section 4.1.3.

Visual Attention: Although no previous work explicitly defines visual attention in reading scenarios, substantial studies demonstrate a strong correlation between eye movement patterns and attentional processing during reading. For instance, the E-Z reader model [66] posits that attention during reading moves from word to word continuously. The serial-processing assumption states that attention is linked to focus changes in text processing [26, 56, 85]. Following these studies, our work describes visual attention during reading by establishing the connection between eye movement patterns and the corresponding while-reading text components, such as words and sentences. Specifically, we define the visual attention state as the collection of eye movement features on each text component. For example, when reading the sentence "*They race to maturity, with the shortest generation time of any vertebrate*", the visual attention for the word "*vertebrate*" consists of fixation duration, reading times, number of fixations, etc. At the sentence level, the visual attention state is defined using the total dwell time, saccade times, etc.

Semantic Attention: We are interested in exploring how the semantic meaning from text assists in estimating the time-series reading states and how they explain these states. From this perspective, it is necessary to have a holistic semantic understanding of while-reading texts. Furthermore, such understanding should cover the semantic meaning of different grain sizes of texts, ranging from single words and sentences to passage levels. This works terms this semantics collection at various granularities as semantic attention. For example, semantic attention can hint at whether the while-reading text components are difficult. These difficult components may be unfamiliar or ambiguous words or sentences with complex syntax, which often delay reading. In this case, appropriately using such semantic meaning regarding the difficult score can provide additional evidence in revealing the current reading state and deliver a reasonable interpretation regarding why the current text components block the reading.

4 SYSTEM DESIGN

This section describes the CASES design. We first detail the CASES network (CASES-Net), a deep neural network for detecting and interpreting ongoing reading states. Then, we describe a real-time reading state estimation and intervention system aiming to boost reading comprehension performance.

4.1 CASES Network

4.1.1 Overall Pipeline. Figure 2 depicts the overall pipeline of the proposed CASES-Net. It consists of four modules: semantic attention extraction (SAE), visual attention extraction (VAE), cross-attention extraction (CAE), and reading state estimation/explanation.

The first step in the CASES-Net pipeline provides a comprehensive semantic understanding of the text before the reading begins. This semantic meaning information compensates for the low-resolution eye-tracking data, thus enabling accurate reading state estimation. Semantic meaning also enables explanations during reading state detection tasks in later pipeline stages. To extract semantic meaning, the system turns on the outward-facing scene camera to obtain the text to be read. The SAE module then runs once on the text. It utilizes NLP techniques to extract the high-resolution semantic features and the inherent linguistic structure from the text, thus facilitating subsequent tasks.

Texts contain rich semantic information, but for better individual reading state estimation, personalized visual attention data are also necessary. To capture it, the VAE module is triggered to obtain the online visual attention features corresponding with text components (e.g., while-gazing words or sentences). More specifically, CASES-Net senses reader eye images to predict gaze sequences using continuous eye-tracking [46, 60]. Then, the VAE module extracts visual attention features from the sequential gaze data. In parallel, the scene camera records time-aligned scene images to help track gaze positions.

Since the obtained semantic meaning of the text and visual attention features are at different spatial resolutions, we propose the CAE module to properly align them. We use words to segment the visual attention features because words are the minimal text units considered in this work, upon which sentences and global context depend.

The TCN-based network estimates the reading states at word and sentence levels, aiming to explore the task-specific features for the assistance of the multi-task output. One feature represents the binary determination of whether a reader has difficulty processing a word; we call this the "word-level task". The second task is hierarchical multi-label classification at the sentence level, which includes (Task I) estimating whether a reader is having sentence-level processing difficulty and if so, (Task II) estimating whether the reader is facing comprehension challenges, the reader's mind is wandering, or both. A multi-task and hierarchical loss function for training guides CASES-Net. We can qualitatively understand the reasons for the predicted reading states by visualizing the learned semantic attention and visual attention features.

The rest of this section explains the technical details of each module.

4.1.2 Semantic Attention Extraction Module. SAE module aims to understand the high-resolution semantic meaning of the document $\mathbf{R}$, ranging from the word level to the document level. There are two primary prerequisites for extracting accurate semantic features: obtaining the while-gazing locations and text contents. The former, i.e., while-gazing locations, can be obtained by using eye tracking and represented as Points of Gaze (PoG) timeseries. Each PoG corresponds to a two-dimensional coordinate in the scene image recorded by the scene camera. Given the locations of PoG, we can easily load the while-gazing text contents because the reading system has already stored all the reading materials in advance. After that, we propose to extract the following three types of semantic features by utilizing various advanced NLP techniques.

(1) Each word in $\mathbf{R}$ is encoded as a 768-dimensional vector by XLNet model [86], which can learn the semantic meaning of the document by processing the whole text passage once. To lower the potential adverse effect incurred by the high dimensionality, we reduce the XLNet features to 64 dimensions via a fully-connected (FC) layer and denote them as $\mathbf{r}^B = \{\mathbf{r}^B_w\}^W_{w=1}$, where $\mathbf{r}^B_w \in \mathbb{R}^{64}$ and W is the total number of words.
(2) To understand the keyword information in the document, we calculate the probability of each word describing the whole document via the YAKE model [6]. The keyword features are denoted as $\mathbf{r}^K = \{r^K_w\}^W_{w=1}$, where $r^K_w \in \mathbb{R}$.

(3) We use word difficulty to assist in the final task of identifying the reading state. Following Franklin et al. [22], we describe the word difficulty using the length of the word, number of syllables, and familiarity scored by the MRC psycholinguistic database [14]. We denote the difficulty of words by $\mathbf{r}^D = \{\mathbf{r}^D_w\}_{w=1}^W$, where $\mathbf{r}^D_w = C(l_w, s_w, f_w) \in \mathbb{R}^3$, C is the concatenation operation, l_w, s_w, and f_w denote the word w's length, syllable number, and familiarity score, respectively.

Finally, each word in the document is represented by the concatenation of the three feature vectors; that is $\mathbf{r}_w = C(\mathbf{r}^B_w, r^K_w, \mathbf{r}^D_w) \in \mathbb{R}^{68}$ $(w = 1, 2, \ldots, W)$. Note that the semantic features regarding more coarse levels (e.g., sentence- and passage- level) can be generalized from that of the word level, as words are inherently structured and semantically connected — a passage consists of multiple sentences and a sentence of multiple words.

4.1.3 Visual Attention Extraction Module. A reliable gaze sequence is a foundation for accurate visual attention feature extraction. However, the raw gaze points are noisy due to difficult-to-avoid human motion and limited eye-tracking resolution. To alleviate this issue, we design a filtering algorithm to smooth the raw gaze points, leveraging their sequential characteristics. More specifically, we first employ an existing eye-tracking technology to estimate the PoGs and record the PoGs sequences as $\mathbf{E} = \{\mathbf{e}_t\}_{t=1}^T$, where T is the total number of timestamps considered. The designed filtering method first uses median filtering to discard outliers due to gaze jitter. Then, we use mean filtering to stabilize the fluctuations of sequential PoGs due to the limited eye-tracking resolution. After filtering, we obtain the smoothed PoGs $\mathbf{E}^* = \{\mathbf{e}^*_t\}_{t=1}^T$. We segment each word and sentence using $\mathbf{E}^*$ and then send them to the next step for visual attention extraction.

The number of PoGs will increase rapidly during reading. To reduce the size of PoGs, experts have engineered a large number of representative features reflecting how people comprehend characters during reading [16, 31, 72, 73] or whether people are disengaged from reading [19, 54]. In this work, we propose to further enrich the engineered visual features. The following features are widely used to describe word-level processing state while reading: fixation duration, number of fixations, and number of repeated word readings. However, we observe that these three features vary not only differ from person to person but also change while reading. Such variation significantly affects estimation performance. The personal variation is usually removed by normalizing personal data [29]; however, the latter while-reading variation is rarely considered. This work introduces local information to tackle the latter problem: every τ seconds, we add the statistical features to describe the mean and the variance of each engineered feature, for $\mathbf{E}^* = \{\mathbf{e}^*_t\}_{t=1}^\tau$, to describe the visual attention for each word. In total, we obtain a 9-dimensional feature for each word. Moreover, we normalize the four sentence-level representative visual features, including dwell time [17], saccade times [20], forward saccade times [59], and backward saccade times [59], using the sentence length, so these features better describe the local variation. Given that we segment M words during τ, the visual feature of each word is represented using $\mathbf{r}^E_w \in \mathbb{R}^{(9+4)}$ $(w = 1, 2, \ldots, M)$. There are nine word-level features and four sentence-level features that are identical to the words in the same sentence.

Lastly, we propose to use the higher-level temporal features of the sequential gaze data, as recent studies have demonstrated the effectiveness of deep neural networks (DNN) on eye movement pattern classification. We adopt the existing feature extractor based on the 1D-CNN with BLSTM backbone [75] (denoted as $\mathcal{N}_{eye}$) to extract 8-dimensional deep features during time duration τ, i.e., $\mathbf{r}^{DNN}_w \in \mathbb{R}^8$ $(w = 1, 2, \ldots, M)$.

4.1.4 Cross-Attention Extraction Module. To facilitate downstream multi-task learning, the CAE module first fuses the two modalities, then explores the commonalities and distinct task-specific information to make predictions at different granularities.

Before fusing the two modalities, we use the following strategy to synchronize them for time alignment. Specifically, for each smoothed PoGs sequence $\mathbf{e}^*_t$, we identify the M words being processed at time t, and concatenate the three features vectors to obtain $\mathbf{f}^w_t = C(\mathbf{r}_w, \mathbf{r}^E_w, \mathbf{r}^{DNN}_w) \in \mathbb{R}^{(68+13+8)}$ as the overall representation of the two modalities. For all other words w' that have not been visually processed till time t, we pad the semantic

attention feature vector $\mathbf{r}_{w'}$ with a zero vector, i.e., $\mathbf{f}_t^{w'} = C(\mathbf{r}_{w'} \in \mathbb{R}^{68}, \mathbf{0} \in \mathbb{R}^{21})$. In this way, the word being processed at time t can be properly described semantically with its corresponding visual attention features. In contrast, the unread words padded with zeros are given less attention.

The CAE module uses a Temporal Convolutional Network (TCN) model, which is capable of capturing temporal dependencies. Specifically, the module uses temporal convolutional filters/kernels to process input sequences. Each filter calculates a weighted average in the time domain, and the parameters of the filters are learned to optimize the objective function. The CAE module has four TCN layers, each of which consists of temporal convolutions, a non-linear ReLU activation function, and a max pooling function or an upsampling function. To achieve efficient learning across different tasks, we divide the filters of the final layer into two types: task-specific filters used for the word- and sentence- level tasks, and task-shared filters for both tasks.

4.1.5 Reading State Estimation and Explanations. After obtaining the cross-attention features, we are ready to detect the reading state of "processing difficulty". We have the following three tasks. (1) Word-level binary-class classification task T^{word}: The word-level features are fed to a fully connected layer (denoted as $\mathcal{N}^{word}$) to predict whether a reader finds the word being processed difficult. Sentence-level and word-level tasks differ. Since we know that mind wandering may co-occur with reading difficulty for a sentence, we formulate the task at the sentence level in the following hierarchical fashion. (2) Sentence-level binary-class classification task $T^{sent,1}$: With the sentence-level features, we first determine whether the reader is in a normal reading state without any processing difficulties using a binary classifier. We use $\mathcal{N}^{sent,1}$ to denote the subnetwork of conducting $T^{sent,1}$. (3) Sentence-level multi-label classification task $T^{sent,2}$: If the reader enters into an abnormal state, the reader can be either mind wandering or processing difficulty, or both; This is a multi-label classification task, where multi labels can be assigned simultaneously; label 1 is mind wandering and label 2 is processing difficulty. We use $\mathcal{N}^{sent,2}$ to denote the subnetwork of conducting $T^{sent,2}$.

Finally, to train the network, we propose the following loss function reflecting the performance of all tasks:

$$\mathcal{L} = \mathcal{L}(T^{word}) + \alpha\mathcal{L}(T^{sent,1}) + \beta\mathcal{L}(T^{sent,2}), \tag{1}$$

where α and β are tradeoff parameters. Binary Cross Entropy (BCE) loss is used for T^{word} and $\mathcal{L}(T^{word})$ is illustrated as follows

$$\mathcal{L}(T^{word}) = -\frac{1}{W}\sum_{w=1}^{W}\left(y_w^{word}\log p_w^{word} + (1-y_w^{word})\log(1-p_w^{word})\right), \tag{2}$$

where W denotes the number of word; y_w^{word} denotes the label of word w, $y_w^{word} = 0$ indicates the reader finds the word w easy, $y_w^{word} = 1$ indicates the reader finds the word w difficult; p_w^{word} is the word-level estimation results given by the network $\mathcal{N}^{word}$.
BCE loss is also used for $T^{sent,1}$ and $\mathcal{L}(T^{sent,1})$ is illustrated as follows

$$\mathcal{L}(T^{sent,1}) = -\frac{1}{S}\sum_{s=1}^{S}\left(y_s^{sent,1}\log p_s^{sent,1} + (1-y_s^{sent,1})\log(1-p_s^{sent,1})\right), \tag{3}$$

where S denotes the number of sentences; $y_s^{sent,1}$ denotes the binary classification label of the sth sentence, $y_s^{sent,1} = 0$ indicates the reader is in a normal reading state for sentence s, $y_s^{sent,1} = 1$ indicates the reader is in an abnormal reading state; $p_s^{sent,1}$ is the sentence-level binary classification estimation results given by the network $\mathcal{N}^{sent,1}$.
For sentences with $y_s^{sent,1} = 1$, to solve the multi-label problem, BCE loss is used for each label separately, and the

loss of $T^{sent,2}$ is illustrated as follows

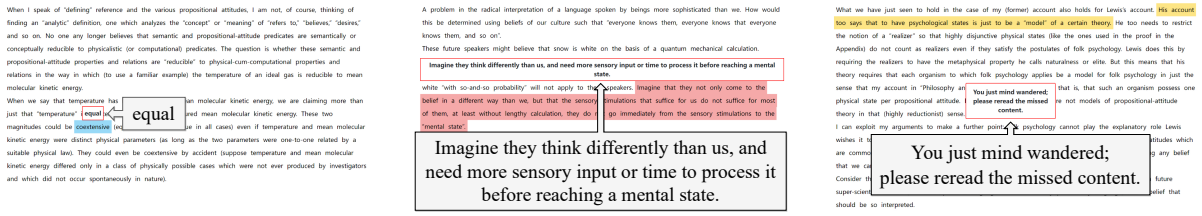
$$\mathcal{L}(T^{sent,2}) = -\frac{1}{\sum_{s=1}^{S}\mathbf{1}(y_s^{sent,1}=1)}\sum_{s=1}^{S}\sum_{l=1}^{L}\mathbf{1}(y_s^{sent,1}=1)\left(y_{s,l}^{sent,2}\log p_{s,l}^{sent,2}+(1-y_{s,l}^{sent,2})\log(1-p_{s,l}^{sent,2})\right), \quad (4)$$

where $L = 2$ denotes the number of labels, i.e., label 1 as mind wandering and label 2 as processing difficulty; $y_{s,l}^{sent,1}$ denotes the supervised information of the lth label for sentence s, $y_{s,l}^{sent,1} = 1$ indicates sentence s has the lth label, $y_{s,l}^{sent,1} = 0$ indicates sentence s does not have the lth label; $\mathbf{1}(\cdot)$ is an indicator function, $\mathbf{1}(y_s^{sent,1} = 1) = 1$ when $y_s^{sent,1} = 1$, $\mathbf{1}(y_s^{sent,1} = 1) = 0$ when $y_s^{sent,1} = 0$; $p_{s,l}^{sent,2}$ is the sentence-level multi-label estimation results given by the network $\mathcal{N}^{sent,2}$.

4.2 EYEReader: A Real-Time Reading State Detection and Intervention System

Our goal is to determine reading state series that influence reading fluency and mitigate the negative effects of reading processing difficulties. To this end, we build a real-time reading state detection and intervention system (called EYEReader) for English language. For the convenience of readers, EYEReader is implemented in the form of a website, enabling cross-platform compatibility.

This section first gives a concrete example to show the key features of EYEReader and how to use it. Then it details the system architecture, along with its operation pipeline. At last, it describes the hardware prototype.



(a) (b) (c)

Fig. 3. Screenshots of three intervention examples. Detection and Interventions: (a) simplifying challenges word at the word level (left), (b) streamlining complex sentences at the sentence level (middle), and (c) giving mind-wandering reminders at the sentence level (right).

4.2.1 Key Features and Operation Process of EYEReader. We first give some key features of EYEReader, and we then use a concrete case to show the automatic detection and intervention process.

Key feature 1: text materials selection. The text materials should contain various topics, as the intervention is anticipated to be text-agnostic. We select 36 reading comprehension materials with diverse topics from an English qualification test to match the participants' reading comprehension ability. Each article has around 450 words on average. Users can log in to the system, select their preferred articles from existing materials, and start reading by simply clicking a button.

Key feature 2: friendly reading interface. Because we have overcome the limited resolution issues when eye-tracking is used during reading scenarios, the interface of text presentation of EYEReader is similar to common computerized reading settings. More specifically, articles are automatically divided into several different pages (around 240 words per page) with a regular line height, approximately single-spaced. We adopt an 18-point default font typeface.

Key feature 3: intervention design. The interventions are designed to help users overcome the three while-reading processing difficulties, i.e., mind wandering, challenging words, and complex sentences, that may lead to a negative impact on their reading comprehension performance. Three interventions are designed for the three difficulties respectively: 1) providing an immediate reminder once mind wandering is detected, which reminds readers to focus on the current reading; 2) simplifying the challenging words; and 3) streamlining the complex sentences. We provide the following three examples to further clarify how the interventions support reading.

(1) **Simplifying challenges words.** Once the system detects that a reader is facing a challenging word, it highlights the word in blue and provides a more comprehensible one in the pop-up window. For example, when a user struggles with "coextensive", the pop-up window offers a more straightforward and easy-to-understand one, "equal". Figure 3 (a) provides a screenshot of this intervention. After receiving the interventions, users can click on the highlighted words to hide the reminders and continue reading.
(2) **Streamlining complex sentences.** The procedure of streamlining complex sentences is similar to that of simplifying challenging words. Differently, the system highlights the complex sentences in red and provides simpler sentences in the pop-up window. For example, for a long and complex sentence, "Imagine that they not only come to the belief in a different way than we, but that the sensory stimulations that suffice for us do not suffice for most of them, at least without lengthy calculation, they do not go immediately from the sensory stimulations to the "mental state"." the system provides a relatively more straightforward version: "Imagine they think differently than us, and need more sensory input or time to process it before reaching a mental state." A screenshot of this type of intervention is depicted in Figure 3 (b). Users can also click on the highlighted sentences to hide the pop-up window and continue reading.
(3) **Giving mind-wandering reminders.** When the system detects that the reader is distracted while reading, it highlights the missed content in yellow and displays a pop-up message in the center of the screen, showing that "You just mind wandered; please reread the missed content." A screenshot of this type of intervention is depicted in Figure 3 (c). The pop-up message automatically fades out after one second.

Participants' eye gazes are calibrated prior to their reading in order to correlate the two cameras equipped in the eyewear. The calibration method follows Pupil Capture[1] [41]. Specifically, during the calibration phase, the participants wear the eyewear and sit in front of the computer, a pupil calibration marker appears on the screen with fixed locations. The participant is instructed to gaze at the maker for approximately two seconds. The same procedure is executed for the other four calibration markers on the screen. In this way, the system would record these positions to correlate the two cameras.

During the reading process, readers wear the prototype eyeglass and sit in front of the computer to read. The trained CASES-Net model is always-on to automatically detect potential abnormal reading states, i.e., whether the user is struggling with difficult words or complex sentences, or their mind is wandering. When abnormal events that affect reading are detected, the system triggers interventions automatically. The text components will be highlighted, and the corresponding treatments will be shown on the right-top of the text content automatically in a pop-up window.

4.2.2 System Architecture. Figure 4 illustrates the overall architecture of EYEReader. We use the Vue.js framework to develop the front-end website, while we choose Django for the back-end of the website, as it is a widely-used Python web framework [5]. Django offers a variety of third-party tools for building communication between the front-end and back-end efficiently following the REST API specification. To store and manage the data on the server, we adopt one of the widely-used open-source database management systems–MySQL [57]. The eyewear and the PC used for reading are connected to the same LAN (Local Area Network). The eyewear is running on

[1] https://docs.pupil-labs.com/core/software/pupil-capture/#calibration

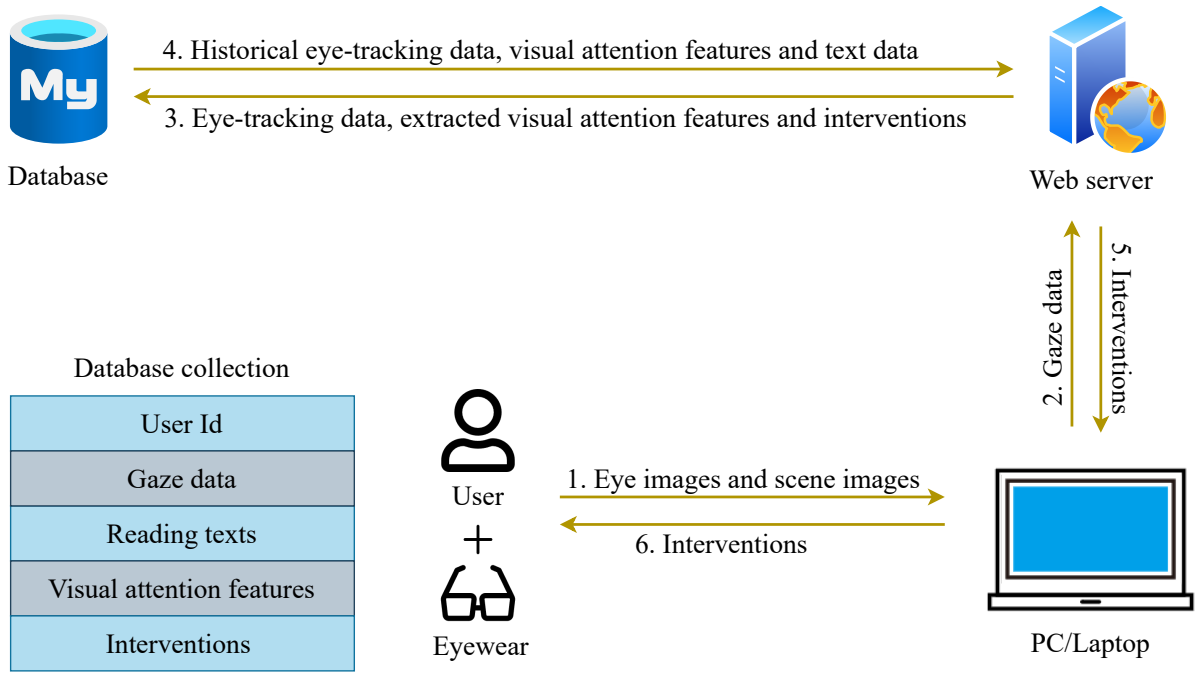


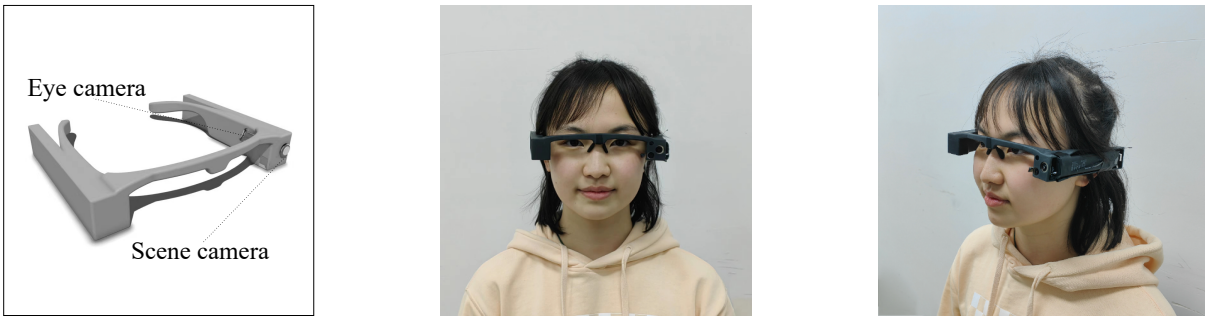
Fig. 4. The architecture of the reading state detection and intervention system.

Android 12. We developed a service app without user interfaces to read the real-time video stream recorded by the two cameras and push the video stream to the PC via the RTSP protocol[2]. On the PC edge, we receive the coming video stream from the eyewear using the RTSP protocol. The received video is then handled by Pupil Capture and Pupil Service provided by Pupil Labs [3].

Next, we describe the overall operation workflow of the built intervention system and show how it provides just-in-time interventions for users encountering reading processing difficulties. There are mainly six steps described below.

- Step 1: During system operation, EYEReader loads the trained CASES-Net from the server when receiving the requests from the front end.
- Step 2: The recorded eye/scene images captured by eyewear are pushed to the user's PC for eye-tracking using the Pupil Capture [41].
- Step 3: The tracked gaze points are sent to the server for further visual attention feature extraction.
- Step 4: The server loads the historical eye-tracking data, visual attention features, and texts to decide when to intervene.
- Step 5: Once processing difficulties are detected, the estimation results are returned to the front end for triggering interventions. The corresponding treatment is shown at the front end to facilitate the current reading.
- Step 6: The current interventions and all other data are saved in Database.

[2] https://en.wikipedia.org/wiki/Real_Time_Streaming_Protocol
[3] https://docs.pupil-labs.com/core/diy/



(a) 3D eyeglass frame (b) Front view (c) Lateral view

Fig. 5. Hardware prototype of CASES eyewear.

4.2.3 Hardware Design. We design prototype eyewear and integrate CASES-Net into the eyewear, as eyewear is a natural way to be used in various reading scenarios.

We presume that the eyewear will be well-migrated to various reading scenarios. Therefore, we adopt a stand-alone scheme to integrate the computing components and power supply into the headset frame. Figure 5 shows the eyewear hardware prototype.

The eye-tracker follows the Pupil [3], and we make slight adjustments to suit our case. More specifically, we use Qualcomm Snapdragon 865 platform directly integrated into the left leg of the eyewear. The eye camera and scene camera modules are replaced with 20 MegaPixels (MP) Samsung S5K3T2 and 64 MP Samsung S5KGW1, respectively. The eye camera is used to record eye videos to perform eye tracking. The scene camera senses scene videos to capture the text being read. We design the 3D eyeglass frame to fit the two cameras into the left leg of the mounting frame. To balance the weight of the headset, the battery is integrated into the right leg of the eyewear.

5 EVALUATIONS

This section describes experiments to evaluate CASES, the cognition-aware eyewear system for estimating reading states. We first detail the experimental setup, data collection, and evaluation measures. Then, we present results and quantify the technical capabilities of CASES. All experimental procedures are approved by the ethical committee at Fudan University.

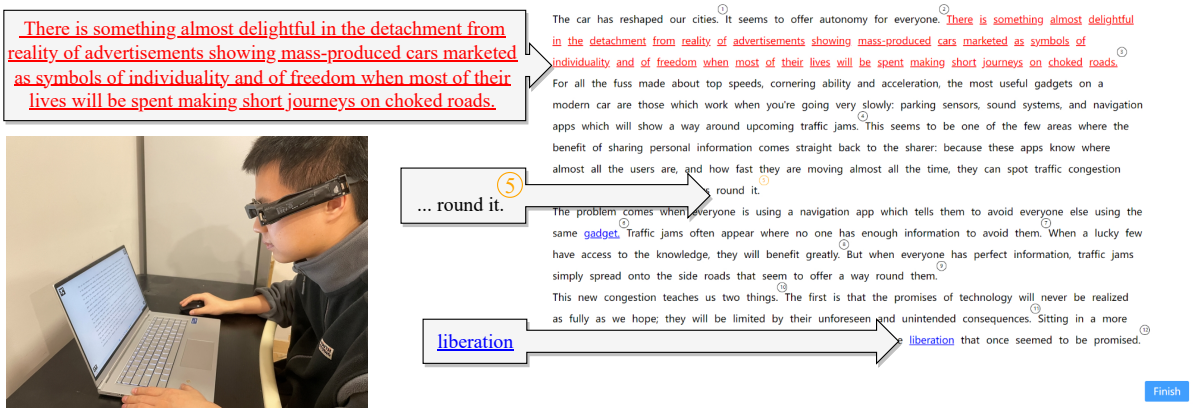
5.1 Evaluation Methodology

5.1.1 Experimental Setup. We recruited 25 participants by posting a questionnaire at the Fudan University campus. We informed the participants about the purpose of our study and the procedure of the experiments before they started. We also gave them the option to withdraw at any time during the experiments. Also, all participants signed an informed consent form. After completing their sessions, the participants received either local currency equivalent to 14 dollars or a thank-you gift worth approximately 14 dollars for their participation.

A summary of the participant demographics follows.

- **Age**: 22–28 years old with an average age of 23.5.
- **Gender ratio**: 19 males (76.0%) and 6 females (24.0%).
- **Native/non-native speaker**: 5 native speakers (20.0%) and 20 non-native ones (80.0%).

As shown in Figure 6 (a), the participant wears eyeglasses and sits in front of the computer to read. While reading, we record videos using the eye camera and time-aligned videos using the scene camera.



(a) A participant is reading

(b) Screenshot from the labeling tool

Fig. 6. The in-lab setting of CASES experimental study.

5.1.2 Text Material Selection. Texts should cover a wide range of subjects so readers can enter multiple reading states. Moreover, each text should be short, allowing participants to read several texts. This study selects 36 articles with the following three subjects:

Subject matter 1: One-minute BBC world news[4]: 10 articles with approximately 300 words per article on average.

Subject matter 2: English qualification tests[5] : 16 articles on reading comprehension materials with approximately 450 words per article on average.

Subject matter 3: Philosophy related [63]: 10 articles with approximately 500 words per article on average.

The first two of these provide challenging words and sentences, respectively. The third may lead to mind wandering. We anticipate that most participants are unfamiliar with the third subject matter, and it is hard to understand the content without prior knowledge. The idea of mundane subject selection to introduce mind wandering follows a recent work [54].

Considering the diverse backgrounds and prior knowledge of various participants, texts should also cover a wide range of subject classes. According to Dewey Decimal Classification (DDC) method [70], we categorize the selected articles into ten subject classes, including "social science", "religion", and eight other subjects. Prior to data collection, we select an approximately equal number of articles from each topic class, except for the philosophy articles.

5.1.3 Dataset. The CASES requires time-aligned eye gaze data and text data (i.e., the words or sentences being read) to detect reading states. In addition, the synchronized data should capture continuous reading, during which users may encounter various reading states. To the best of our knowledge, there are no publically available datasets suitable for our problem. Therefore, we develop an online system to collect data meeting our requirements.

[4]https://www.bbc.com/news
[5]https://cet.neea.edu.cn/

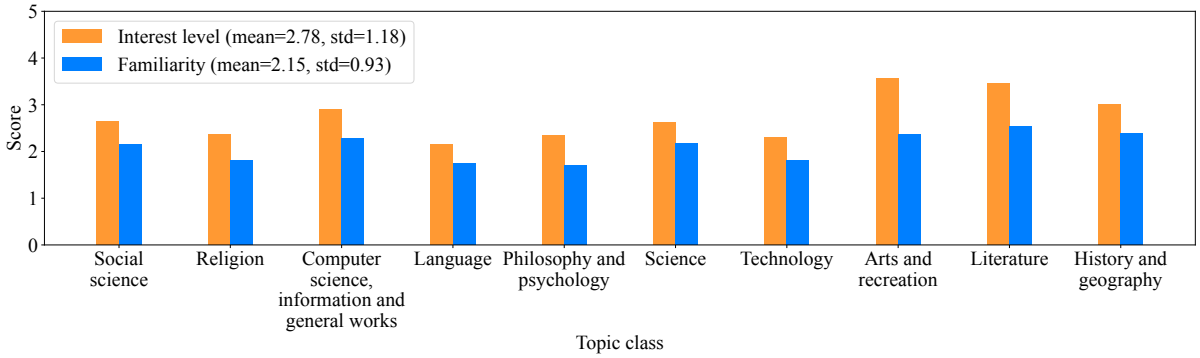


Fig. 7. Score of interest level and familiarity for each topic class.

Next, we detail the data collection procedure. The collected dataset is available online at[6] to facilitate the relevant research.

(1) Data Collection. The above articles are randomly assigned to each participant. Specifically, we randomly select articles from each topic for the participants to ensure that they cover all three subject matters. This design allows most participants to encounter numerous reading states. Each article is divided into pages. There are around 240 words per page in single-spaced 18-point typeface. Then, we verbally instruct participants on how to use the data-collection system, such as navigating to the next/previous page. Finally, each participant reads the texts. Reading one article takes approximately six minutes.

(2) Ground-Truth Labeling. After completing an article, the participant is immediately instructed to label their reading states. We developed a labeling tool with a GUI window to accelerate labeling. Participants can review each page of the article. They mark the challenging words and sentences they do not comprehend on each page using single and double clicks, respectively. We also provide a button at the top right of each sentence for users to mark whether their minds wandered when reading it. The annotated words and sentences are highlighted in different colors so users can quickly double-check their annotations. Figure 6 (b) provides a screenshot from the labeling tool. Annotating one article takes around three minutes. In total, the data collection process, including the annotation collection, took us approximately fourteen days.

(3) Dataset Statistics. The collected dataset is randomly split into training (80%) and test (20%) sets per participant/article. The total numbers of labels for "word-level processing difficulties"/"sentence-level processing difficulties"/"mind wandering" are 1005/244/200.

We survey the participants' interest level and familiarity with the ten topics to further verify the fairness of the selected topics, i.e., we expect that the interest level and familiarity are evenly distributed across all topics. Using the Likert scale[7], we ask participants to score their interest level and familiarity with the articles they read. The scale ranges from 1 to 5; a higher score indicates that a participant is more familiar with or more interested in the article. As shown in Figure 7, participants gave roughly similar interest scores (mean = 2.78, std = 1.18) and familiarity (mean = 2.15, std = 0.93) on the ten topic classes, indicating that the ten topic classes have covered individual participants evenly. The means of interest level and familiarity of all topics are around 2.5, suggesting that the topics are intermediate to participants.

[6] https://github.com/MemX-Research/CASES
[7] https://en.wikipedia.org/wiki/Likert_scale

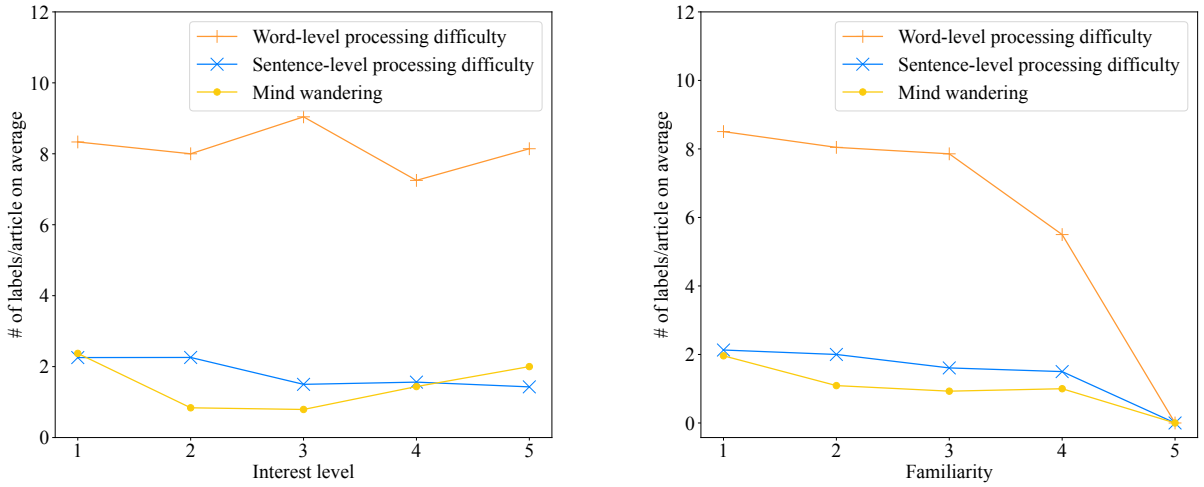


Fig. 8. Number of labels per article on average at each interest level and familiarity.

We also visualize the distribution of the average number of labels per article at various levels of interest and familiarity in Figure 8. As can be seen in Figure 8 (left), readers give approximately the same number of labels per article under each interest level. Also, Figure 8 (right) shows that the number of labels per article decreases as familiarity increases. This is in line with our intuition, as participants often give more labels for their unfamiliar articles.

5.1.4 Evaluation Metrics. Because our framework is hierarchical and multi-task, we need to adopt appropriate measures to evaluate each task. The first task is a binary classification of whether a reader is facing difficulty processing a word. We evaluate its performance using accuracy and the receiver operating characteristics (ROC) curve. The second task is hierarchical multi-label classification at the sentence level, which includes sentence-level Task I and Task II. As Task I is a binary classification of whether a reader has sentence-level processing difficulty, we also use accuracy and ROC curve to evaluate it. On the other hand, Task II is multi-labeled. Following previous work [88], we use the multilabel-based macro-averaging metric, i.e., averaged-accuracy and ROC curve, to evaluate it.

5.1.5 Baseline Methods. We conduct ablation studies to evaluate CASES, as there is no prior work solving the problem addressed in this work, thus making direct comparisons with prior work infeasible. We use the following three baseline methods for evaluation.

(1) Visual: Previous studies have demonstrated that some reading states, such as mind wandering, can be identified using gaze-relevant features [19, 54], which are closely related to our work. To validate whether the gaze-relevant features are sufficient for reading state recognition at multiple text element granularities (words and sentences), this work uses a baseline method leveraging 13 gaze-relevant features (9 word-level features and 4 sentence-level features described in Section 4.1.3) to identify the state while reading. We use the support vector machine (SVM) method to conduct the three classification tasks: word-level task, sentence-level Task I, and sentence-level Task II. This work adopts SVM as it has been successfully applied to various classification tasks [78], and is one of the widely used methods in similar tasks [20, 54]. For simplicity, we refer to this method as *Visual*.

(2) Visual+: Eye movement patterns are good indicators for reading state recognition. Inspired by prior work [75] that leverages deep neural network (DNN) to achieve accurate eye movement pattern identification, we use the 8-dimensional higher-level temporal features extracted from a deep neural network (1D-CNN with BLSTM [75]) to improve the accuracy of reading state estimation. To make a fair comparison, the extracted deep features are concatenated with the aforementioned 13 gaze-relevant features and sent to the CAE module removing the semantic attention feature concatenation part to estimate reading states. This baseline method is an improved version of the Visual method called *Visual+*.

(3) NLP: Visual and Visual+ identify reading states based solely on visual attention features. To verify the classification performance based on the semantic context of texts, we designed this baseline method, dubbed *NLP*. As in the Visual+ method, we first extract semantic features using the SAE module and then send the extracted features to the CAE module without the visual attention feature concatenation part to infer reading states.

5.2 Results

5.2.1 Overall Performance. Figure 9 shows the reading state recognition performance of our method and three baseline methods. CASES achieves the best performance among all methods. Compared with the Visual method, i.e., conventional eye-tracking only, CASES improves the accuracy by 6.85%, 8.55%, 20.90% for the word-level task and the sentence-level Task I and Task II. Furthermore, compared with the baseline method Visual+ and NLP, CASES has superior reading state estimation. For example, the sentence-level Task II detection accuracy of CASES is 86.64% while it is 79.15% or lower for the baseline methods. We conclude that using context derived from text improves reading state estimation.

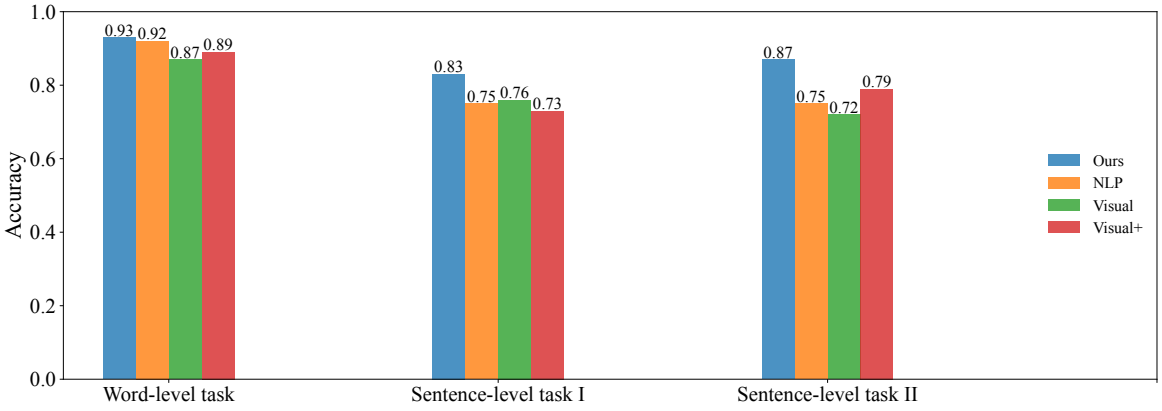
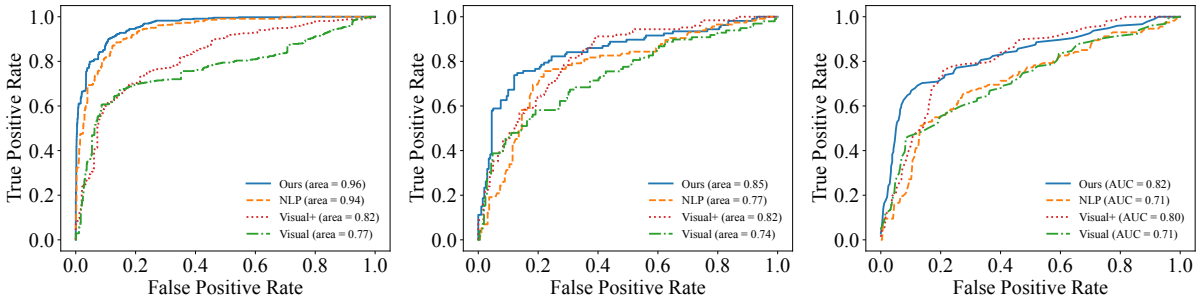


Fig. 9. Reading state classification accuracy for CASES and the baseline methods.

We plot the ROC of different methods. Figures 10a, 10b, and 10c demonstrate that CASES outperforms the baseline methods in Area Under the Curve (AUC), which is one of the most widely used performance measures in classification or retrieval problems.

The following section further explains why CASES outperforms the baseline methods and how it offers semantic explanations of the predicted reading states.



(a) Word-level task (b) Sentence-level Task I (c) Sentence-level Task II

Fig. 10. ROC for CASES and the baseline methods.

6 PILOT STUDY

This work aims to study progression through cognitive states while reading to assist our understanding of the reading process. To this end, we have conducted in-field pilot studies using CASES, the proposed system, for three and a half months. This section first summarizes the initial findings around our designed two RQ and hypotheses using CASES. Then, it demonstrates the capability of EYEReader to make helpful real-time interventions when reading difficulties are encountered. Finally, it revisits the two RQ, describes the limitations of our system, and indicates possible extensions of this work.

6.1 The Procedure of the Pilot Study

We recruited thirteen volunteers to participate in the pilot study from Fudan University. The average age is 23.9 years (SD=1.6, min=22, max=28), with n=4 (30.8%) female and n=9 (69.2%) males. There are 10 non-native speakers (76.9%) and 3 native ones (23.1%). The non-native speakers reported that they had passed Level 6 of the College English Test (CET6) in our country[8], and the native speakers were college-level students at our University.

During the pilot study, participants wore the prototype eyeglasses, sat in front of the computer, and logged in to the development website to read. The participants either took the eyeglasses with them and used the eyeglasses whenever they would like to do the experiments or came to our laboratory for the experiments. Participants were encouraged to use the system whenever they read, as reasonable observations require the prolonged engagement of participants.

The pilot study lasted three and a half months and consisted of two stages. During the first stage, we required participants to label the words and sentences they encountered difficulty processing, and these labels were treated as ground truth. Based on the qualitative evaluation [68], we examined the labeled data point by point at different granularities around the designed RQ and hypotheses. Then, we made several findings on how people read at different granularities, i.e., single words and sentences, and summarized the following six patterns to discuss. The second stage focused on applying EYEReader in practice. At the end of the pilot study, each participant completed a survey of their opinions on the usability and value of EYEReader. Finally, we confirmed the proposed hypotheses.

6.2 Key Observations

6.2.1 Observations at Word Level. This section presents three observations on how users read at the single-word level.

[8] https://en.wikipedia.org/wiki/College_English_Test

Observation I: Users comprehend the lexical meanings of words by directing their gazes more frequently toward material they find difficult to process. When users encounter difficulty processing a word, they usually gaze at it longer and more times than typical. This observation is consistent with prior evidence about the process of comprehending single words during reading [16, 53]. Figure 11 illustrates one example of this observation, where participant P6 has difficulty comprehending the meaning of "mitigate" and "debris". P6 fixates "mitigate" (fixation label 13) and "debris" (with fixation label 19) for a long time and reads them more than two times. In particular, P6 has the longest fixation duration on the word "debris" and has the most reading times on the word "debris" and "mitigate".

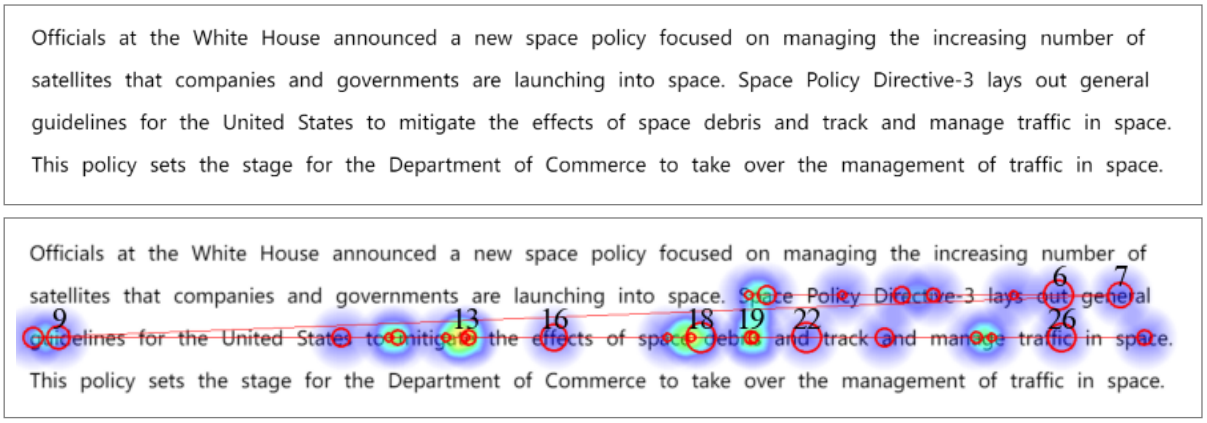


Fig. 11. Visualization of visual attention for P6 reading a sentence. Each circle represents a fixation point. The larger the area of the circle, the longer the fixation duration. The circle number denotes the timestamp of the fixation time. *Top:* Raw text; *Bottom:* Text with filtered point of gazes

Figure 12 provides another example of this observation for a native participant. Participant P12* (∗ indicates native user hereinafter) is facing challenging words "counterbalanced" (with fixation label 10) and "sketch" (fixation labels 19 and 21). Under the text context presented in Figure 12, we observe that P12* has the most prolonged fixation duration on words "sketch" and "counterbalanced".

Observation II: When a user encounters difficulty processing a word, the user first directs their gaze to the word and then to other words to examine the semantic context. Readers generally avoid breaking their chain of thinking by stopping when a difficult word is encountered, especially when the word does not affect their understanding of the text. However, when readers consider a difficult word highly topic-relevant or meaningful for subsequent text comprehension, they tend to interrupt their reading and attempt to deduce the semantic meaning of the word from its semantic context. This observation differs from a previous study [16], and our next observation complements it.

Observation III: When users examine the semantic context of a difficult-to-process word, they gather semantic clues by shifting their gazes to different locations even when considering the same difficult word, from the same text, under similar reading conditions. Readers typically attempt to find an appropriate location in the text to help comprehend the current difficult-to-process word. The text at the location should reveal the relevant information about the difficult word. Also, that location varies from person to person, depending on their current cognitive states about the context.

Figure 13 shows the proportion of the three above observations for each participant by summarizing their past experienced processing difficult words. We observe that the ten non-native participants experience Observation I

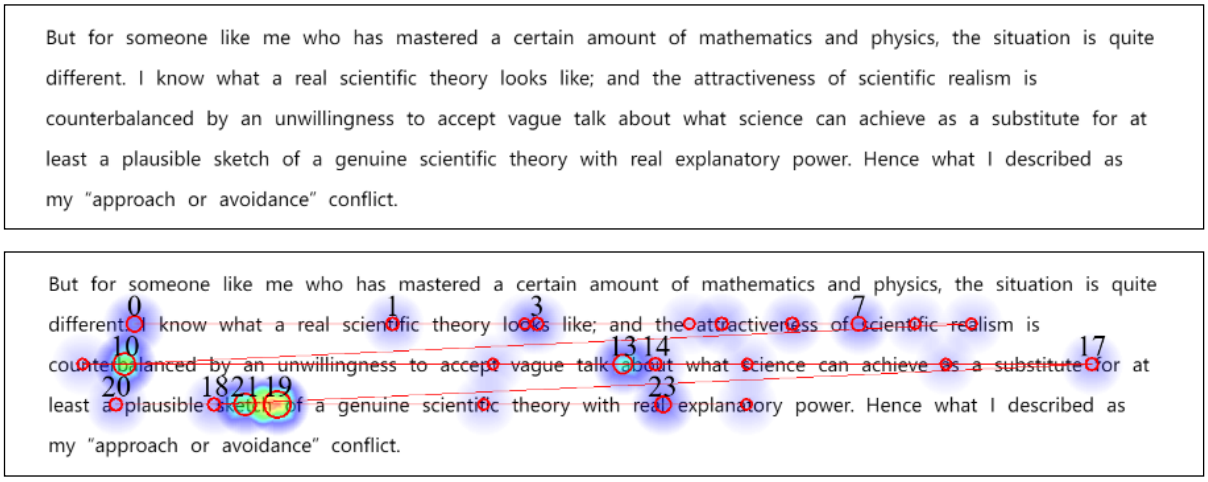


Fig. 12. Visualization of visual attention for P12* reading a sentence. *Top:* Raw text; *Bottom:* Text with filtered point of gazes.

in most cases (around 87.47% cases on average), and they fall into Observation II & Observation III in fewer times, i.e., around 12.53% on average. In contrast, the native participants experience Observation II & Observation III in most cases (around 60.67% cases on average), and they fall into Observation I fewer times, i.e., around 39.33% on average. This aligns with our intuition, as we anticipate that native readers are more adept at leveraging the context cues from texts to help their reading comprehension.

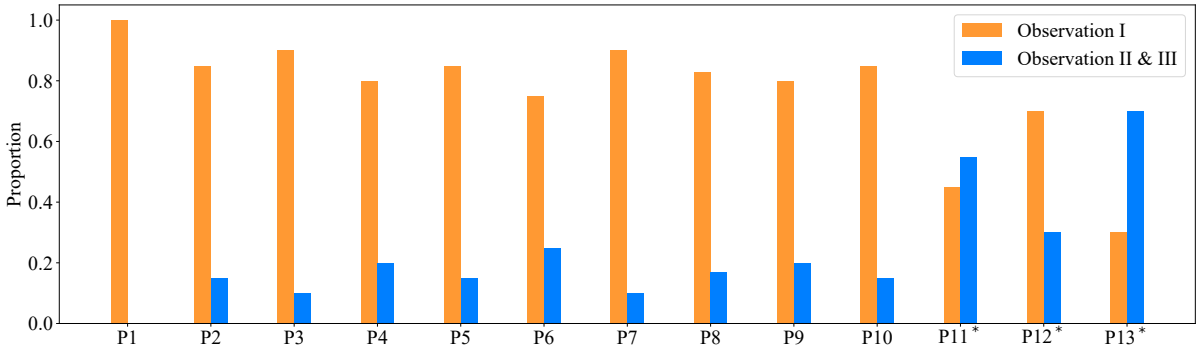


Fig. 13. Profile of word-level comprehension failures for thirteen readers. ∗ indicates native speakers.

Figure 14 shows an exemplary case to provide further insights on Observation II and Observation III. Here two readers, P2 and P5, face the same reading difficulty in comprehending the word "liberation" when they read the same sentence from the same article. We can see that the two participants first direct their visual attention to the target word, "liberation" where the fixation labels are 14 and 13 for P2 and P5, respectively. They then shift their gazes. Participant P2 gazes back at the previously read word, "pleasant", while Participant P5 gazes forward to the word, "promised". Both of these words are semantically relevant to the difficult word, "liberation", as shown in Figure 14 (top row).

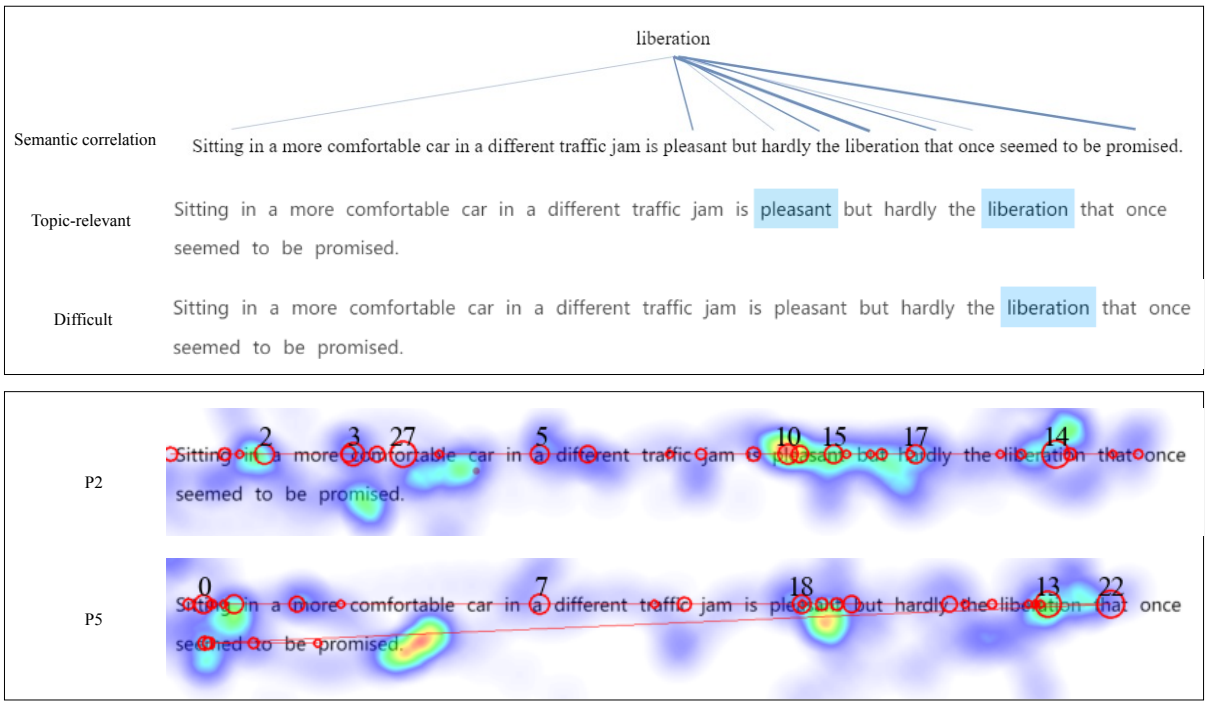


Fig. 14. An example where two non-native speakers struggle to comprehend the word "liberation".

Figure 15 provides an example for two native speakers, P12* and P13*. They face the same reading difficulty in comprehending the challenging word "realism". Clearly, P12* and P13* first direct their gaze to "realism" with fixation labels 14 and 11, respectively, and then shift their gazes. P12* gazes forward to an antonym word "antirealism". Differently, P13* gazes back at the already-read word "internal", a modifier of the target word.

6.2.2 Observations at Sentence Level. This section focuses on two modes of comprehending sentences: interpretive (semantic) and structural (syntactic).

Observation IV: People incrementally comprehend the semantics of a sentence as they read each word, while with different gaze time series. Figure 16 and Figure 17 show the inter-reader differences in gaze time series when reading the same sentence for non-native and native speakers, respectively. As shown in Figure 16, native speaker P1 focuses on the first parts of sentences (with more fixation, labels 0–12) while P4 focuses on other parts of sentences (fixation labels 9–11). A similar phenomenon can be seen in Figure 17, two native speakers, P13* and P11*, read the sentence sequentially but with different visual focuses. P13* focuses on the first parts of the sentence, e.g., with more distinct locations of focus, while P11* focuses on other parts of sentences.

Observation V: Readers enter the "rereading" or "reanalysis" state at different times when having difficulty with the same sentence. Figure 18 depicts such an example for two non-native speakers, P8 and P3. Both users face challenges comprehending the sentence "*The authors, who ..... in recent years.*" P8 backtracks 3–4 words (with a fixation label starting from 28) when reading the middle of the sentence and then continues reading the sentence; while P3 rereads the sentence from the beginning when reading the middle of the sentence (fixation label 12).

Figure 19 depicts such an example for two native speakers, P11* and P13*. Both users face challenges comprehending the sentence "*As countless boards and ..... and overall performance.*" We observe that P11* rereads the

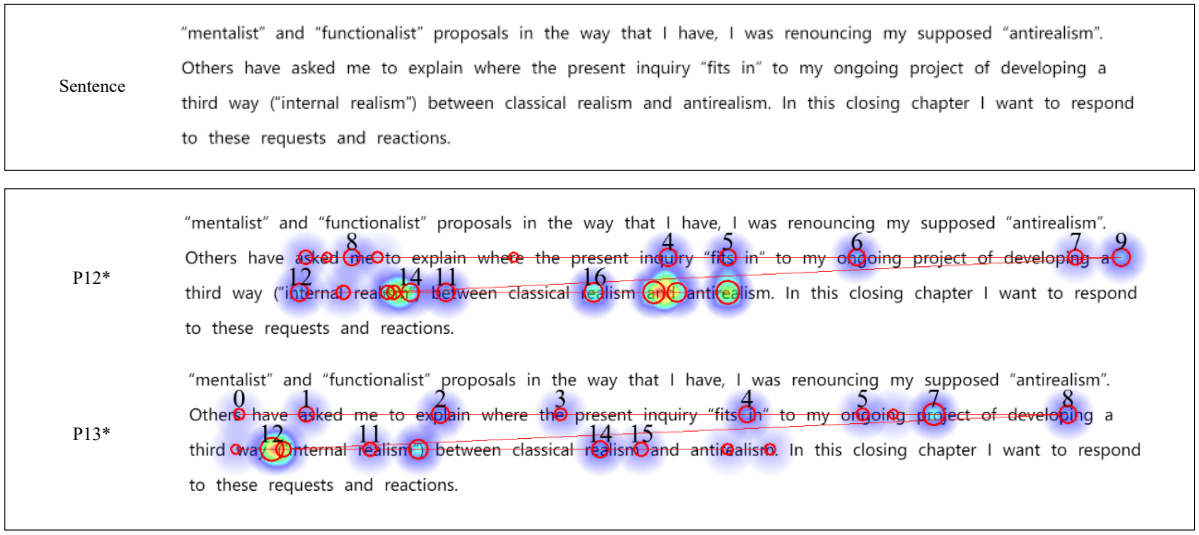


Fig. 15. An example where two native speakers have difficulty understanding the word "realism".

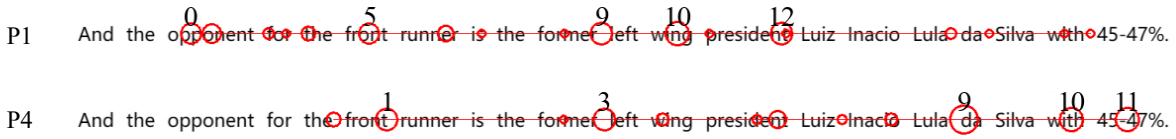


Fig. 16. An example of two non-native speakers reading the same sentence with different visual attention.

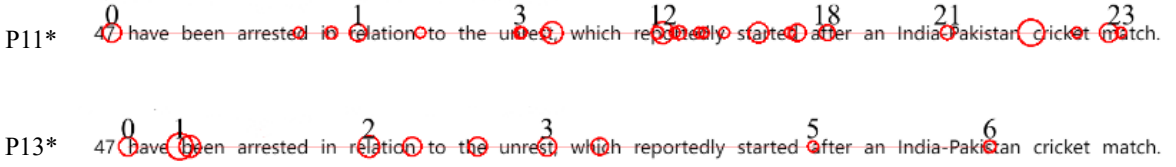


Fig. 17. An example of two native speakers reading the same sentence with different visual attention.

sentence (fixation label 23) right after completing the first-pass reading (fixation label 22); while P13* rereads the sentence from the beginning of the sentence (fixation label 6) when finishing the middle of the sentence (fixation label 5).

Observation VI: Different people "reread" the same sentence with different reading states. Figure 20 shows two non-native speakers reading the same sentence twice. P1 gets distracted (i.e., enters the mind wandering state) during the first reading of the sentence (typical fixation labels 2, 6, and 14); therefore, P1 spends more time and has more fixations on the sentence in the second reading (fixation labels 21, 24, 26, 28) than in the first pass. In contrast, P4 spends more time reading the sentence the first time (fixation labels 3, 17, and 18) but quickly skims it the second time (fixation labels 26 and 37).

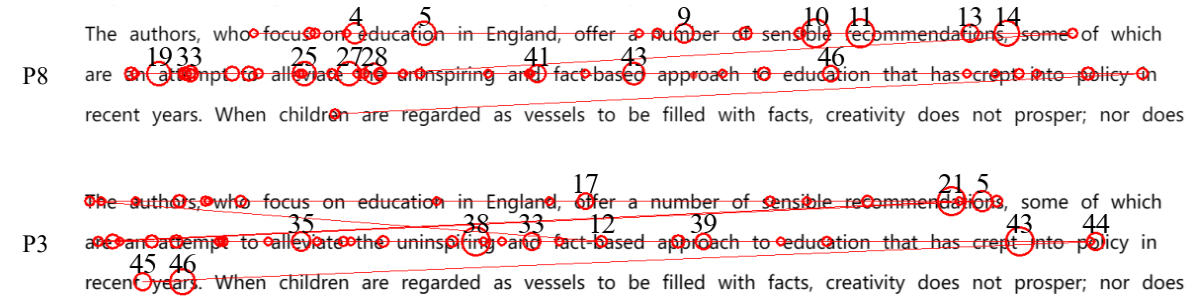


Fig. 18. An example of two non-native speakers having different "reread" behaviors when encountering comprehension difficulties on the same sentence.

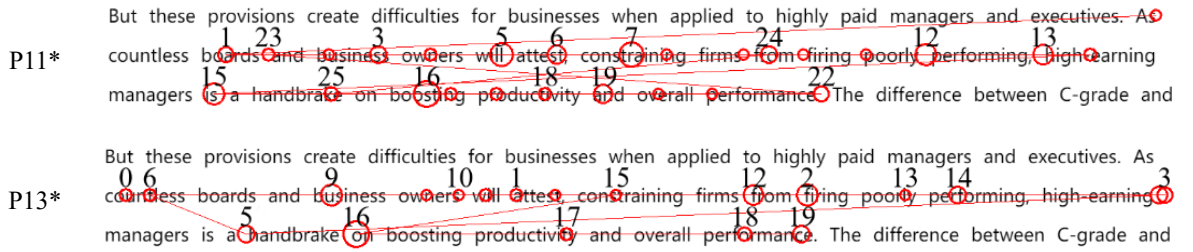


Fig. 19. An example of two native speakers having different "rereading" behaviors when encountering comprehension difficulties on the same sentence.

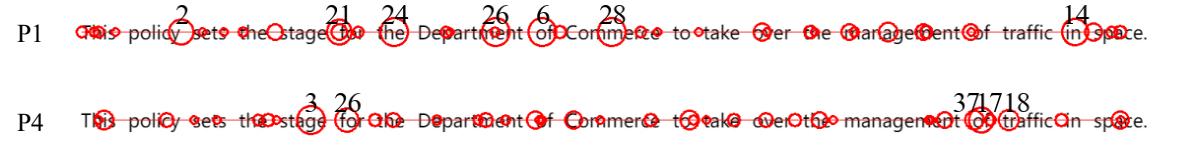


Fig. 20. An example of two non-native speakers "rereading" the same sentence with different reading states. P1 gets distracted during the first reading and then rereads the sentence. P4 also reads the sentence twice, but without any comprehension difficulties, i.e., he is in a normal state of comprehension.

Figure 21 shows two native users, P11* and P12*, reading the same sentence twice. They label their reading states as sentence-level processing difficulty and mind wandering. P11* spends more time and more fixations when reading the sentence in the first pass (typical fixation labels 1, 5, and 18) than in the second pass (typical fixation labels 26, 29, and 36). Also, P11* rereads the sentence after completing the next sentence (fixation label 25). Differently, P12* gets distracted during the first pass of the sentence (typical fixation labels 5-15); therefore, P12* rereads the sentence with more time and has more fixations on the sentence in the second pass (fixation labels 18, 21, 26 and 27).

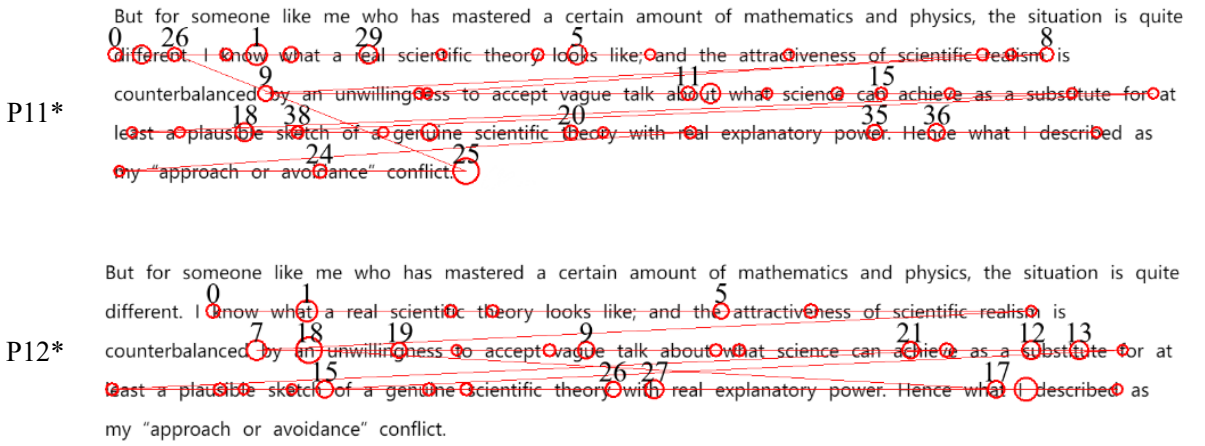


Fig. 21. An example of two native speakers "reread" the same sentence with different reading states. P11* encounters processing difficulty with the sentence, and then P11* rereads the sentence; P12* gets distracted during the first reading and then rereads the sentence.

6.3 Evaluation of EYEReader in Practice

Section 5 shows that CASES can accurately detect reading states. This section evaluates the ability of EYEReader to improve reading comprehension by identifying reading states implying processing difficulties and making real-time interventions.

To make quantitative assessment of EYEReader, we define the improvement of reading comprehension performance (called reading gain for short in this work) as $(s_{past} - s_{present})/s_{past}$, where $s_{present}$ and s_{past} denote the number of challenging words or sentences at present and in the past, respectively. The higher the $(s_{past} - s_{present})/s_{past}$, the higher the reading comprehension improvement. As shown in Figure 22, all thirteen participants show non-negative reading gains, and the majority of them have significant improvements, with most achieving a reading gain of at least 0.20. That means EYEReader is effective in helping users to overcome unfamiliar words and complex sentences.

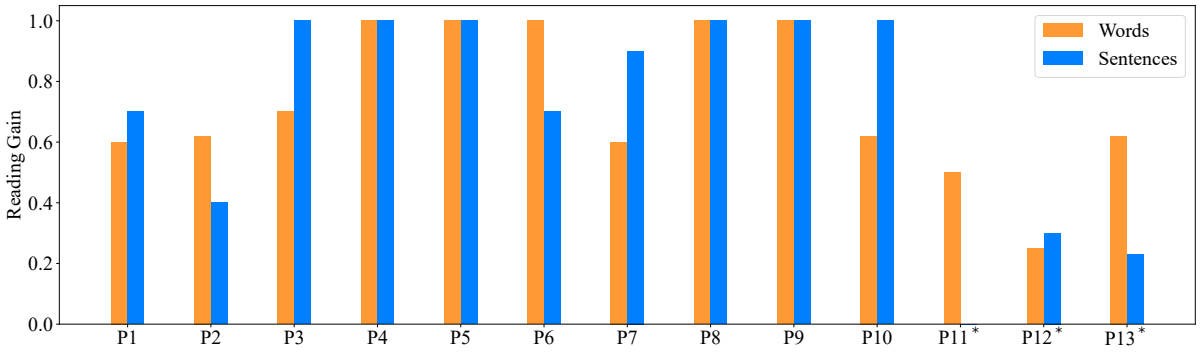


Fig. 22. Profile of reading gain for 13 participants in pilot studies. ∗ indicates native speakers.

6.4 Feedback from Participants

We designed several open-ended questionnaires to qualitatively evaluate EYEReader. Thirteen questionnaires were sent to participants, twelve of which were returned. Among them, 10/12 of the participants positively commented on word-level intervention. They believe that fine-grained intervention at the word level can precisely pinpoint the reading difficulties they are experiencing. Nine out of the twelve participants reported that the sentence-level intervention was helpful. In particular, when facing challenging sentences with complex syntactic structures, it was difficult to comprehend the sentences even though they were familiar with all the words. In this case, EYEReader helped them overcome this reading difficulty by highlighting and explaining the sentence. In addition, 9/12 of the participants found EYEReader valuable in reminding them when their minds wandered; these participants stated that they usually do not realize when distracted. Timely reminders can make their reading more focused and efficient.

Furthermore, we collect participants' opinions regarding whether the eyewear hardware will negatively affect their reading process. Through conducting a questionnaire, we asked the participants in the pilot study to give scores for the comfortable level of hardware on a scale of 1-5; the corresponding description is listed in Table 1. A total of 13 questionnaires were sent out, and 12 were recalled. Statistical results show that participants generally think the hardware has a media or negligible impact on their reading process (mean = 3.33, std = 0.75).

In addition, once the trained CASES-Net is applied to practice, we design the proper system intervention so as not to break the chain of thoughts of users. That is, the intervention can help readers avoid interrupting reading due to the encountered processing difficulties that lead them to seek help from other means [32], such as a dictionary. Therefore, we design the intervention process with minimal interaction cost and encourage readers to focus on the current reading. To assess the impact of the intervention on reading, we also conducted a questionnaire to collect readers' opinions regarding the user-friendliness of intervention interaction. Similarly, we asked the participants in the pilot studies to give scores on a scale of 1-5; the corresponding description is listed in Table 2. Statistical results show that most participants think the intervention process is user-friendly (mean = 3.92, std = 0.76).

Table 1. Rating scale of the comfort level with regards to how the hardware affects the reading process.

Score	Description
1	Severe impact
2	Significant impact
3	Neutral
4	Negligible impact
5	No impact totally

Table 2. Rating scale regarding whether the intervention interaction is user-friendly.

Score	Description
1	Very unfriendly
2	Unfriendly
3	Neutral
4	Friendly
5	Very friendly

6.5 Discussion and Future Work

CASES has the goal of accurately estimating and providing semantic explanations of reading states over time, which can facilitate the scientific study of reading by enabling a deeper understanding of the cognitive processes involved in learning to read, disentangling the complex combination of cognitive skills and their impact on reading fluency, and measuring the efficacy of methods for teaching reading and beneficial reading habits.

Next, we first revisit the proposed research questions and hypotheses. Then, we briefly discuss the potential future works that will improve CASES.

6.5.1 Revisiting Research Questions and Hypotheses. We confirm the hypotheses for the two presented RQ based on the results and observations, which we detail below.

RQ1: Do readers in the same reading states show different visual attention distributions on the reading text?

Confirming hypothesis 1: Readers in the same reading state do show varying visual attention histories. As inter-person variation, i.e., individual difference, is ubiquitous, the visual attention histories of readers in the same reading states indeed differ from each other, which can be found from Observation II, Observation III, Observation IV, Observation V, and Observation VI.

RQ2: When readers are in the same reading states, e.g., encountering difficulty progressing, how does reader visual attention interact with semantic cues in the text?

Conforming hypothesis 2: When readers encounter the same processing difficulties, they shift their visual attention to the surrounding text to fetch contextual semantic cues. In other words, when readers' reading progress is blocked, easy text that is semantically related to complex text also receives more visual attention and cognitive effort, which can be found from Observation II and Observation III.

6.5.2 *Discussion and Future Work.*

(1) Science of reading. This work investigates the human cognitive reading process by exploring the complementarity of eye movements and text. However, it is also important to integrate illustration information to understand how people read. A recent study has shown that text-diagram instructions can improve reading comprehension [36]. Thus, our future work aims to exploit semantic information, including text and illustrations, and integrate them with eye movements to investigate the reading cognitive progress.

In addition, we aim to investigate more reading states that might provide a complete picture of the reading cognitive progress. In addition to determining the reading states at the word and sentence levels, it would be valuable to measure how people read at the entire passage level. This could deepen our understanding of how people summarize and reflect on learned knowledge during reading.

(2) Interactive Reading System. Our system is still an early-stage prototype. A longer user study would enable the collection of more data and user feedback to improve the interactive design and user experience. This could help us to build a mature reading assistance system that contributes to educational applications, HCI studies, etc.

(3) Reading Contexts. We would like to emphasize that we presume that the system will be well-migrated to various reading scenarios, and therefore, we use the eyeglasses form to study reading. We believe wearing eyeglasses to read is a portable way in numerous reading contexts, including computerized reading and physical reading (e.g., reading newspapers). However, since our eyeglasses are still in the early prototype stage, in this work, we did not experimentally cover all the scenarios. The system presented in this work is currently used in a computerized-reading context, as reading using electronic devices has become common in our modern lives and has been widely studied by a large body of researchers [13, 30, 54]. We are aware that investigating the physical-reading context is also important, and we are interested in applying our eyeglasses to investigate the reading (cognitive) states under this context in our future work.

(4) Brain-Sensing Methods in Reading. In addition to eye-tracking in reading, brain-sensing via electroencephalograph can determine the level of cognitive workload under different rapid serial visual presentation settings, as demonstrated in [45]. It can be utilized to determine the cognitive workload or attention of texts at different granularity levels. However, this has to be done with the eye movement data jointly to accurately locate the positions of text being read and allow fine-grained analysis on processing difficulty of words. We believe that it is a direction that is worth exploring in the future to further improve the performance of our system.

7 CONCLUSIONS

This work presents CASES, a cognition-aware smart eyewear system that automatically recognizes reading (cognitive) state timeseries using eye tracking and text semantic context. We conduct ablation studies to demonstrate that CASES significantly improves the accuracy of reading state recognition over the conventional approach

that relies only on eye tracking. Furthermore, in-field studies enable several observations about how individual reading state timeseries are related to text semantic context at different granularities. The ability to track semantic context cues enables better understanding of progressive reading states. We embody CASES in an interactive reading assistant system that provides just-in-time interventions when users encounter reading difficulties. Several months of deployment demonstrate the benefits of the system in promoting self-awareness of cognitive processes while reading and improving reading comprehension performance. We envision that CASES will be of use in the scientific study of reading, cognition, and human-computer interfaces.

REFERENCES

[1] Ugo Ballenghein, Johanna K Kaakinen, Geoffrey Tissier, and Thierry Baccino. 2020. Cognitive engagement during reading on digital tablet: Evidence from concurrent recordings of postural and eye movements. *Quarterly Journal of Experimental Psychology* 73, 11 (2020), 1820–1829.

[2] Gregory S Berns, Kristina Blaine, Michael J Prietula, and Brandon E Pye. 2013. Short-and long-term effects of a novel on connectivity in the brain. *Brain connectivity* 3, 6 (2013), 590–600.

[3] Yevgeni Berzak, Chie Nakamura, Amelia Smith, Emily Weng, Boris Katz, Suzanne Flynn, and Roger Levy. 2022. CELER: A 365-participant corpus of eye movements in L1 and L2 English reading. *Open Mind* 6 (2022), 41–50.

[4] Stephen Bottos and Balakumar Balasingam. 2019. Tracking the Progression of Reading Through Eye-gaze Measurements. In *22th International Conference on Information Fusion*. IEEE, 1–8.

[5] Carl Burch. 2010. Django, a web framework using python: Tutorial presentation. *Journal of Computing Sciences in Colleges* 25, 5 (2010), 154–155.

[6] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences* 509 (2020), 257–289.

[7] Jon W Carr, Valentina N Pescuma, Michele Furlan, Maria Ktori, and Davide Crepaldi. 2022. Algorithms for the automated correction of vertical drift in eye-tracking data. *Behavior Research Methods* 54, 1 (2022), 287–310.

[8] Benjamin T Carter and Steven G Luke. 2020. Best practices in eye tracking research. *International Journal of Psychophysiology* 155 (2020), 49–62.

[9] Yuhu Chang, Yingying Zhao, Mingzhi Dong, Yujiang Wang, Yutian Lu, Qin Lv, Robert P Dick, Tun Lu, Ning Gu, and Li Shang. 2021. MemX: An attention-aware smart eyewear system for personalized moment auto-capture. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–23.

[10] Shiwei Cheng, Zhiqiang Sun, Lingyun Sun, Kirsten Yee, and Anind K. Dey. 2015. Gaze-Based Annotations for Reading Comprehension. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1569–1572.

[11] Michelene TH Chi and Ruth Wylie. 2014. The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist* 49, 4 (2014), 219–243.

[12] KR1442 Chowdhary and KR Chowdhary. 2020. Natural language processing. *Fundamentals of artificial intelligence* (2020), 603–649.

[13] Virginia Clinton. 2019. Reading from paper compared to screens: A systematic review and meta-analysis. *Journal of research in reading* 42, 2 (2019), 288–325.

[14] Max Coltheart. 1981. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A* 33, 4 (1981), 497–505.

[15] Anne Cunningham and Keith Stanovich. 2003. Reading Can Make You Smarter!. *Principal* 83, 2 (2003), 34–39.

[16] Pablo Delgado and Ladislao Salmerón. 2022. Cognitive Effort in Text Processing and Reading Comprehension in Print and on Tablet: An Eye-Tracking Study. *Discourse Processes* 59, 4 (2022), 237–274.

[17] Ekaterina Denkova, Jason S Nomi, Lucina Q Uddin, and Amishi P Jha. 2019. Dynamic brain network configurations during rest and an attention task with frequent occurrence of mind wandering. *Human brain mapping* 40, 15 (2019), 4564–4576.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 4171–4186.

[19] Sidney K D'Mello, Caitlin Mills, Robert Bixler, and Nigel Bosch. 2017. Zone out No More: Mitigating Mind Wandering during Computerized Reading. In *Proceedings of the 10th International Conference on Educational Data Mining*. International Educational Data Mining Society (IEDMS).

[20] Myrthe Faber, Robert Bixler, and Sidney K D'Mello. 2018. An automated behavioral measure of mind wandering during computerized reading. *Behavior Research Methods* 50, 1 (2018), 134–150.

[21] Aisha Farid, Muhammad Ishtiaq, and Muhammad Sabboor Hussain. 2020. A Review of Effective Reading Strategies to Teach Text Comprehension to Adult English Language Learners. *Global Language Review* 5, 3 (2020), 77–88.

[22] Michael S Franklin, Jonathan Smallwood, and Jonathan W Schooler. 2011. Catching the mind in flight: Using behavioral indices to detect mindless reading in real time. *Psychonomic bulletin & review* 18, 5 (2011), 992–997.
[23] Edward Gibson and Evelina Fedorenko. 2013. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes* 28, 1-2 (2013), 88–124.
[24] Amber Gove and Peter Cvelich. 2011. Early reading: Igniting education for all. A report by the early grade learning community of practice. *RTI International* (2011).
[25] Malcolm Haynes and Thad Starner. 2018. Effects of lateral eye displacement on comfort while reading from a video display terminal. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–17.
[26] John M Henderson and Fernanda Ferreira. 1993. Eye movement control during reading: fixation measures reflect foveal but not parafoveal processing difficulty. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 47, 2 (1993), 201.
[27] Riku Higashimura, Andrew Vargo, Motoi Iwata, and Koichi Kise. 2022. Helping Mobile Learners Know Unknown Words through their Reading Behavior. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM, 249:1–249:5.
[28] Jin Huang, Chun Yu, Yuntao Wang, Yuhang Zhao, Siqi Liu, Chou Mo, Jie Liu, Lie Zhang, and Yuanchun Shi. 2014. FOCUS: enhancing children's engagement in reading by using contextual BCI training sessions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1905–1908.
[29] Michael Xuelin Huang, Tiffany CK Kwok, Grace Ngai, Hong Va Leong, and Stephen CF Chan. 2014. Building a self-learning eye gaze model from user interaction data. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 1017–1020.
[30] Stephen Hutt, Kristina Krasich, Caitlin Mills, Nigel Bosch, Shelby White, James R Brockmole, and Sidney K D'Mello. 2019. Automated gaze-based mind wandering detection during computerized learning in classrooms. *User Modeling and User-Adapted Interaction* 29, 4 (2019), 821–867.
[31] Jukka Hyönä and Richard K Olson. 1995. Eye fixation patterns among dyslexic and normal readers: effects of word length and word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21, 6 (1995), 1430.
[32] Aulikki Hyrskykari. 2006. Utilizing eye movements: Overcoming inaccuracy while tracking the focus of attention during reading. *Comput. Hum. Behav.* 22, 4 (2006), 657–671.
[33] Aulikki Hyrskykari, Päivi Majaranta, Antti Aaltonen, and Kari-Jouko Räihä. 2000. Design issues of iDICT: a gaze-assisted translation aid. In *Proceedings of the Eye Tracking Research & Application Symposium, ETRA 2000, Palm Beach Gardens, Florida, USA, November 6-8, 2000*. ACM, 9–14.
[34] Md. Rabiul Islam, Shuji Sakamoto, Yoshihiro Yamada, Andrew W. Vargo, Motoi Iwata, Masakazu Iwamura, and Koichi Kise. 2021. Self-supervised Learning for Reading Activity Classification. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3 (2021), 105:1–105:22.
[35] Ariel N James, Scott H Fraundorf, Eun-Kyung Lee, and Duane G Watson. 2018. Individual differences in syntactic processing: Is there evidence for reader-text interactions? *Journal of memory and language* 102 (2018), 155–181.
[36] Yu-Cin Jian. 2021. The immediate and delayed effects of text–diagram reading instruction on reading comprehension and learning processes: evidence from eye movements. *Reading and Writing* 34, 3 (2021), 727–752.
[37] Yu-Cin Jian. 2022. Influence of science text reading difficulty and hands-on manipulation on science learning: An eye-tracking study. *Journal of Research in Science Teaching* 59, 3 (2022), 358–382.
[38] Philip C. Jackson Jr. 2018. Natural language in the Common Model of Cognition. In *Postproceedings of the 9th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2018 (Ninth Annual Meeting of the BICA Society), August 22-24, 2018, Prague, Czech Republic (Procedia Computer Science, Vol. 145)*. Elsevier, 699–709.
[39] Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: from eye fixations to comprehension. *Psychological review* 87, 4 (1980), 329.
[40] Yuki Kamide and Anuenue Kukona. 2018. The influence of globally ungrammatical local syntactic constraints on real-time sentence comprehension: Evidence from the visual world paradigm and reading. *Cognitive Science* 42, 8 (2018), 2976–2998.
[41] Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *The 2014 ACM Conference on Ubiquitous Computing, UbiComp '14 Adjunct, Seattle, WA, USA - September 13 - 17, 2014*. ACM, 1151–1160.
[42] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2022. Natural language processing: State of the art, current trends and challenges. *Multimedia tools and applications* 82, 3 (2022), 3713–3744.
[43] Jumpei Kobayashi and Toshio Kawashima. 2019. Paragraph-based faded text facilitates reading comprehension. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 162.
[44] Arnout W Koornneef and Jos JA Van Berkum. 2006. On the use of verb-based implicit causality in sentence comprehension: Evidence from self-paced reading and eye tracking. *Journal of Memory and Language* 54, 4 (2006), 445–465.
[45] Thomas Kosch, Albrecht Schmidt, Simon Thanheiser, and Lewis L Chuang. 2020. One does not simply RSVP: mental workload to select speed reading parameters using electroencephalography. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–13.

[46] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. 2016. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE Computer Society, 2176–2184.
[47] Richard L Lewis, Shravan Vasishth, and Julie A Van Dyke. 2006. Computational principles of working memory in sentence comprehension. *Trends in cognitive sciences* 10, 10 (2006), 447–454.
[48] Tal Linzen. 2018. What can linguistics and deep learning contribute to each other? *arXiv preprint arXiv:1809.04179* (2018).
[49] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics* 4 (2016), 521–535.
[50] Steven G Luke and Kiel Christianson. 2018. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods* (2018), 826–833.
[51] Robert A Mason and Marcel Adam Just. 2007. Lexical ambiguity in sentence comprehension. *Brain research* 1146 (2007), 115–127.
[52] Joseph E Michaelis and Bilge Mutlu. 2017. Someone to read with: Design of and experiences with an in-home learning companion robot for reading. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. ACM, 301–312.
[53] Brian W Miller. 2015. Using reading times and eye-movements to measure cognitive engagement. *Educational psychologist* 50, 1 (2015), 31–42.
[54] Caitlin Mills, Julie Gregg, Robert Bixler, and Sidney K D'Mello. 2021. Eye-mind reader: An intelligent reading interface that promotes long-term comprehension by detecting and responding to mind wandering. *Human–Computer Interaction* 36, 4 (2021), 306–332.
[55] Simon Moe and Michael Wright. 2013. Can accessible digital formats improve reading skills, habits and educational level for dyslectic youngsters?. In *International Conference on Universal Access in Human-Computer Interaction (Lecture Notes in Computer Science, Vol. 8011)*. Springer, 203–212.
[56] Robert E Morrison. 1984. Manipulation of stimulus onset delay in reading: evidence for parallel programming of saccades. *Journal of Experimental psychology: Human Perception and performance* 10, 5 (1984), 667.
[57] AB MySQL. 2001. MySQL.
[58] Rustam Nazurty, Nurullaningsih Priyanto, Sarmandan Anggia Pratiwi, and Amirul Mukminin. 2019. Learning strategies in reading: The case of Indonesian language education student teachers. *Universal Journal of Educational Research* (2019), 2536–2543.
[59] Pablo Oyarzo, David D Preiss, and Diego Cosmelli. 2022. Attentional and meta-cognitive processes underlying mind wandering episodes during continuous naturalistic reading are associated with specific changes in eye behavior. *Psychophysiology* 59, 4 (2022), e13994.
[60] Alexandra Papoutsaki, Patsorn Sangkloy, James Laskey, Nediyana Daskalova, Jeff Huang, and James Hays. 2016. WebGazer: Scalable Webcam Eye Tracking Using User Interactions. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. IJCAI/AAAI Press, 3839–3845.
[61] Ellie Pavlick. 2022. Semantic Structure in Deep Learning. *Annual Review of Linguistics* 8, 1 (2022), 447–471.
[62] Charles A. Perfetti. 2000. Comprehending written language: a blueprint of the reader. In *The Neurocognition of Language*. Oxford University Press, 167–208.
[63] Hilary Putnam. 1988. *Representation and reality*. MIT press.
[64] Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin* 124, 3 (1998), 372–422.
[65] Erik D. Reichle. 2006. Computational models of eye-movement control during reading: Theories of the "eye–mind" link. *Cognitive Systems Research* 7 (2006), 2–3.
[66] Erik D. Reichle, Keith Rayner, and Alexander Pollatsek. 2003. The E-Z Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences* 26, 4 (2003), 445–476.
[67] Erik D. Reichle, Andrew E. Reineberg, and Jonathan W. Schooler. 2010. Eye Movements During Mindless Reading. *Psychological Science* 21, 9 (2010), 1300–1310.
[68] Margarete Sandelowski. 1995. Qualitative analysis: What it is and how to begin. *Research in nursing & health* 18, 4 (1995), 371–375.
[69] Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences* 118, 45 (2021), e2105646118.
[70] Mona L. Scott. 1998. Dewey decimal classification. *Libraries Unlimited* (1998).
[71] Prafull Sharma and Yingbo Li. 2019. Self-supervised contextual keyword and keyphrase retrieval with self-labelling. (2019).
[72] John L. Sibert, Mehmet Gokturk, and Robert A. Lavine. 2000. The reading assistant: eye gaze triggered auditory prompting for reading remediation. In *Proceedings of the 13th Annual ACM Symposium on User Interface Software and Technology*. 101–107.
[73] Subroto Singha. 2021. *Gaze Based Mind Wandering Detection Using Deep Learning*. Ph. D. Dissertation. Texas A&M University-Commerce.
[74] Matthew S Starr and Keith Rayner. 2001. Eye movements during reading: Some current controversies. *Trends in cognitive sciences* 5, 4 (2001), 156–163.
[75] Mikhail Startsev, Ioannis Agtzidis, and Michael Dorr. 2019. 1D CNN with BLSTM for automated classification of fixations, saccades, and smooth pursuits. *Behavior Research Methods* 51, 2 (2019), 556–572.

[76] Sawitri Suwanaroa. 2021. Factors and Problems Affecting Reading Comprehension of Undergraduate Students. *International Journal of Linguistics, Literature and Translation* 4, 12 (2021), 15–29.
[77] Amos van Gelderen, Rob Schoonen, Reinoud D. Stoel, C.M. de Glopper, and Jan Hulstijn. 2007. Development of adolescent reading comprehension in language 1 and language 2: A longitudinal analysis of constituent components. *Journal of Educational Psychology* 99, 3 (2007), 477–491.
[78] Vladimir N. Vapnik. 1999. *The nature of statistical learning theory*. Springer science & business media.
[79] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*. 5998–6008.
[80] Shaun Wallace, Zoya Bylinskii, Jonathan Dobres, Bernard Kerr, Sam Berlow, Rick Treitman, Nirmal Kumawat, Kathleen Arpin, Dave B. Miller, Jeff Huang, and Ben D. Sawyer. 2022. Towards Individuated Reading Experiences: Different Fonts Increase Reading Speed for Different Individuals. *ACM Transactions on Computer-Human Interaction* 29, 4 (2022), 38:1–38:56.
[81] Shang Wang and Erin Walker. 2021. Providing Adaptive Feedback in Concept Mapping to Improve Reading Comprehension. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 677:1–677:11.
[82] Edward W. Wlotko and Kara D. Federmeier. 2015. Time for prediction? The effect of presentation rate on predictive sentence comprehension during word-by-word reading. *Cortex* 68 (2015), 20–32.
[83] Fang-Ying Yang. 2017. Examining the reasoning of conflicting science information from the information processing perspective—an eye movement analysis. *Journal of Research in Science Teaching* 54, 10 (2017), 1347–1372.
[84] Li Yang and Tam Shu Sim. 2017. Metacognitive awareness of reading strategies among EFL high school students in China. *AJELP: Asian Journal of English Language and Pedagogy* 5 (2017), 34–45.
[85] Shun-nan Yang. 2006. An oculomotor-based model of eye movements in reading: The competition/interaction model. *Cognitive Systems Research* 7, 1 (2006), 56–69.
[86] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems*. 5754–5764.
[87] Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-Aware BERT for Language Understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 9628–9635.
[88] Yingying Zhao, Yuhu Chang, Yutian Lu, Yujiang Wang, Mingzhi Dong, Qin Lv, Robert P Dick, Fan Yang, Tun Lu, Ning Gu, et al. 2022. Do smart glasses dream of sentimental visions? Deep emotionship analysis for eyewear devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–29.