Don't Wanna Miss a Thing: Gaze-Aware Implicit Interventions for Distraction Recovery in Foreign-Language Videos

MOHAMMED AHMED, Ontario Tech University, Canada
BENEDICT LEUNG, Ontario Tech University, Canada
MARIANA SHIMABUKURO, Ontario Tech University, Canada
CHRISTOPHER COLLINS, Ontario Tech University, Canada

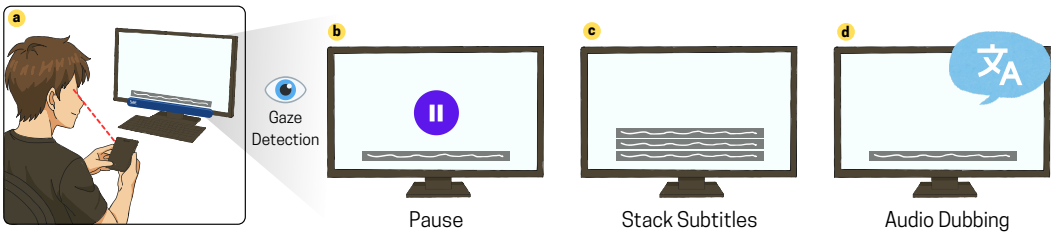

Fig. 1. An overview of the gaze-aware interventions during video-viewing in non-native languages. (a) Gaze detection identifies the viewer's attention away from the screen. Three intervention methods are introduced to resurface missing content: (b) Playback is paused to prevent content loss. (c) Subtitles are stacked to preserve missed dialogue. (d) Audio dubbing provides catch-up narration.

Watching subtitled videos in a foreign language demands sustained visual attention, which can put viewers at risk of missing content due to distraction, such as checking notifications. In this work, we introduced a gaze-aware video player that adapts playback to support attention recovery. We evaluated three gaze-aware techniques: adaptive pausing, stacked subtitles, and audio language switching (dubbing). In a comparative study with 24 participants, we evaluated these techniques against a standard video player with subtitles. While adaptive pausing improved task performance and reduced distractions, stacked subtitles helped recover reading but occasionally slowed faster readers. The benefit of dubbing was limited, resulting in additional cognitive load during the process. Ultimately, all gaze-aware interventions outperformed the standard video player. This work highlights gaze-adaptive systems that seamlessly support attention recovery into everyday viewing experiences.

CCS Concepts: • **Human-centered computing** → **Interaction design**; **Interaction devices**; *User studies.*

Additional Key Words and Phrases: eye-tracking, implicit intervention, distraction recovery

ACM Reference Format:
Mohammed Ahmed, Benedict Leung, Mariana Shimabukuro, and Christopher Collins. 2026. Don't Wanna Miss a Thing: Gaze-Aware Implicit Interventions for Distraction Recovery in Foreign-Language Videos. *Proc. ACM Hum.-Comput. Interact.* 10, 3, Article ETRA015 (May 2026), 18 pages. https://doi.org/10.1145/3806029

Authors' Contact Information: Mohammed Ahmed, Ontario Tech University, Oshawa, Canada, mohammed.ahmed2@ontariotechu.net; Benedict Leung, Ontario Tech University, Oshawa, Canada, benedict.leung1@ontariotechu.net; Mariana Shimabukuro, Ontario Tech University, Oshawa, Canada, mariana.shimabukuro@ontariotechu.ca; Christopher Collins, Ontario Tech University, Oshawa, Canada, christopher.collins@ontariotechu.ca.

ACM 2573-0142/2026/5-ARTETRA015
https://doi.org/10.1145/3806029

1 Introduction

Video has become one of the most popular formats for learning, entertainment, and communication. For many viewers, especially when engaging with content in an unfamiliar language, comprehension relies heavily on subtitles, which provide a text translation of the spoken dialogue. However, following a video with subtitles requires sustained attention and substantial cognitive effort [Alghamdi et al. 2022; Borghini and Hazan 2018; Peng and Wang 2016]. Yet video viewing is often interrupted. Distractions from notifications, multitasking, or mind-wandering can break focus, producing gaps in understanding [Bunce et al. 2010; Lee et al. 2021; Lindquist and McLean 2011; Xiao and Wang 2016, 2017]. When distractions occur, subtitles continue to advance, leaving non-fluent viewers with missing context. Prior eye-tracking research shows subtitle reading is highly consistent, with non-fluent viewers relying heavily on subtitles for comprehension [Elisa Perego and Mosconi 2010]. When attention lapses, this dependency makes recovery especially difficult. Unlike text, which affords selective re-reading, video forces viewers to rely on manual recovery strategies such as scrubbing through the timeline or re-watching entire segments [Pavel et al. 2014]. These strategies are imprecise and time-consuming, and for viewers navigating a foreign language, identifying exactly what was missed or regaining narrative context can be especially challenging.

Gaze-aware interfaces offer a promising direction to address these shortcomings. Eye-tracking provides a reliable indicator of attention [Hyrskykari et al. 2005; Kurzhals et al. 2017]. Gaze-based systems can predict learners' attention and engagement [Bidwell and Fuchs 2011; Hutt et al. 2017; Veliyath et al. 2019], and these estimates can enhance learning performance, especially in video-based learning, where attention and engagement are closely linked to comprehension and outcomes [Arakawa and Yakura 2021; Baker et al. 2010; D'Mello et al. 2012].

Building on this foundation, we investigate how implicit gaze-based interaction can support distraction recovery during foreign-language video viewing, a scenario where comprehension depends strongly on subtitles. We focus on short-term gaze-triggered recovery from momentary lapses rather than long-term learning outcomes, establishing a basis for adaptive media experiences. Our gaze-aware video player monitors viewers' visual behaviour and adapts playback and subtitle presentation accordingly. When attention shifts away, the system triggers one of three interventions (stacked subtitles, adaptive pausing, or audio language switching) to help viewers regain missed information with minimal disruption to viewing flow (Figure 1).

A user study with 24 participants evaluated three gaze-aware intervention techniques for subtitled video viewing: adaptive pausing, stacked subtitles, and dubbing. Participants performed distraction and comprehension tasks while watching videos in another language. The study investigates how these interventions support attention management and recovery from distraction during subtitle-dependent viewing and examines how factors such as cognitive load and user preferences shape their use.

2 Related Work

Research on enhancing viewing experiences during video playback spans attention prediction, subtitle design, and gaze-aware interaction. Previous studies have investigated how viewers process subtitles, how to detect and reduce attention lapses, and how gaze can be used to adapt interfaces. We review these areas to identify gaps motivating gaze-aware interventions for non-fluent viewers.

2.1 Attention-Based Interaction Techniques

Watching videos in a foreign language demands sustained attention to subtitles [Muñoz 2017], making non-fluent viewers particularly susceptible to missed information when distractions occur. Prior work has proposed interaction techniques that monitor non-verbal cues to infer attention levels

and dynamically adapt playback [Arakawa and Yakura 2021; Bidwell and Fuchs 2011; D'Mello et al. 2012; Sharma et al. 2016; Thomas and Jayagopi 2017; Veliyath et al. 2019; Zaletelj and Košir 2017]. In educational contexts, such attention-aware systems have been shown to improve engagement and learning outcomes [Baker et al. 2010; D'Mello et al. 2012].

For example, Mindless Attractor [Arakawa and Yakura 2021] estimates engagement from head pose and subtly perturbs the audio to redirect attention without overt disruption. Other systems automatically pause lecture videos whenever the viewer is taking notes [Nguyen and Liu 2016]. However, most attention-aware systems have been designed for educational settings, and little is known about how these strategies generalize to everyday video viewing. Research on attention-aware interventions in foreign language videos is scarce, even though the interaction between audio and subtitle is known to influence comprehension and viewer comfort [Abu-Rayyash et al. 2024; Liao et al. 2022]. Additionally, the user experience of attention-aware interventions in non-instructional settings, such as entertainment or news videos, remains underexplored. Understanding these contexts could inform attention-aware systems that assist non-fluent viewers while preserving natural and uninterrupted viewing.

2.2 Subtitle Processing

Subtitles are one of the most common aids to support comprehension in unfamiliar languages. They provide a textual representation of spoken dialogue and can enhance understanding while reducing cognitive load [Baranowska 2020; Chan et al. 2022; Kruger et al. 2013]. For non-fluent viewers, subtitles play an especially critical role, enabling word recognition, vocabulary acquisition, and content recall when the spoken language is unfamiliar [Markham et al. 2001; Mitterer and McQueen 2009; Perego et al. 2010].

Eye-tracking research has provided fine-grained insights into the processing of subtitles. Fixations can indicate attention and processing effort [David-John et al. 2021; Pickering et al. 2004]. In a subtitled video, fixation counts and durations reveal reading effort, language proficiency, and comprehension. Beginners tend to skip subtitles less often than advanced learners [Muñoz 2017]. Unlike fixations on general video content, which vary with personal strategies and visual complexity [Elisa Perego and Mosconi 2010; Zheng et al. 2019], subtitle reading is highly consistent: viewers begin reading as soon as subtitles appear, even without prior training, typically without reducing attention to the image [Elisa Perego and Mosconi 2010; Kruger and Steyn 2013; Negi and Mitra 2020]. This tendency is even stronger when the soundtrack is in an unfamiliar language, and subtitles carry essential information [d'Ydewalle and De Bruycker 2007; Elisa Perego and Mosconi 2010].

However, subtitles assume uninterrupted attention. When distractions occur, they continue to advance in sync with dialogue, leaving non-fluent viewers with missing information that can be difficult to reconstruct. As they depend heavily on subtitles, even brief lapses can lead to comprehension breakdowns. Addressing this issue requires adaptive subtitles and audio presentations that can detect and respond to lapses in real time, supporting recovery and maintaining engagement.

2.3 Gaze-Aware Interfaces for Videos

Gaze-aware interfaces dynamically adjust content presentation based on where users look, offering opportunities to enhance accessibility and interactivity in video playback [Matulewski et al. 2018; Nguyen and Liu 2016; Ward et al. 2016]. For example, previous research has explored gaze-adaptive subtitle placement to avoid obscuring important visuals [Kurzhals et al. 2020]. In immersive video systems, playback pauses when users fixate on subtitles and resumes when attention shifts away [Duchowski et al. 2025].

Designing gaze-aware video interfaces involves trade-offs between comprehension support and viewing continuity. Techniques such as gaze-triggered pausing [Duchowski et al. 2025; Nguyen and Liu 2016], replaying missed segments, or switching audio to the viewer's native language [Liao et al. 2022] can aid comprehension but risk disrupting immersion. Effective designs must balance detection accuracy [Bidwell and Fuchs 2011; Veliyath et al. 2019], multimodal synchronization [Abu-Rayyash et al. 2024], and user acceptance across diverse contexts.

Overall, gaze-aware interfaces show promise for supporting non-fluent viewers through dynamic adjustments of subtitles and audio. Such approaches can mitigate information loss during lapses while maintaining the viewing flow. Yet, few studies have examined how to integrate gaze-driven interventions into everyday video viewing, leaving open opportunities to explore seamless, adaptive systems for foreign-language comprehension.

3 Gaze Aware Video Player

This study explores three implicit interaction techniques: attention-based video playback, attention-based dubbing, and subtitle stacking. The first two techniques assess user attention based on whether the gaze is off the screen, while subtitle stacking requires more precise eye-tracking, assuming attention has lapsed once the subtitles are no longer being read. These techniques aim to reduce cognitive load. We employed a traditional video player layout, featuring subtitles at the bottom of the screen and offering foreign language audio tracks alongside English subtitles.

3.1 Attention-based Video Playback

This technique builds on early work on attentive interfaces, which used gaze presence to automatically pause and resume video playback based on viewer attention [Vertegaal 2002]. This intervention technique uses the user's gaze position to determine whether their attention is on the screen. Video playback pauses immediately when the user's gaze is no longer present on the screen and automatically resumes when the gaze returns. This enables the user to recover from distractions without missing a second of their video. While similar gaze-contingent playback mechanisms exist [Duchowski et al. 2025; Nguyen and Liu 2016], we apply this technique to subtitle-mediated comprehension, where missing brief dialogue can impair understanding. This intervention method assumes that users can keep up with the displayed subtitles and allows their gaze to move freely between the subtitle region and the video content.

3.2 Attention-based Dubbing

This method also uses the user's current gaze position to determine whether their attention is focused on the screen. When the user's gaze is no longer detected on the screen, video playback will not pause; instead, it will seamlessly switch to the active language track that the user is familiar with (i.e. English). This approach builds on existing multi-language audio track switching, but differs in that the switch is triggered implicitly based on real-time gaze-detected attention rather than explicit user input. Where a secondary audio track is not present, a text-to-speech audio dubbing is created using the process outlined below. This allows users to listen to the video content while completing their task, even if they are distracted.

Audio-dubbing Process. An audio dubbing process was used for videos that lacked an English-language audio track. We developed a subtitle parser in Python to feed individual subtitle lines into a speech synthesis engine. Initially, we aimed to generate the English audio track while matching the speaker's voice by re-implementing voice cloning techniques. Although the audio produced was impressive in replicating the speaker's voice for multilingual speech synthesis, the overall audio quality was poor due to background noise, and it failed to deliver consistent results.

Fig. 2. Example of gaze-aware subtitle stacking. Subtitles appear at the bottom of the screen, displaying the previously spoken dialogue. The video pauses when three subtitles are stacked on top of each other.

To address this issue, we used background noise suppression with Audacity. However, due to inconsistencies in the quality of the generated speech, we ultimately decided to use Amazon's Polly text-to-speech engine. After synthesizing and saving the subtitles for each video individually, we compared the audio length to the start and end times of each subtitle line. We then adjusted the audio length, stretching or compressing it to match the subtitles without altering the pitch, using a phase vocoder. This ensured that the synthesized audio's speech rate aligned perfectly with the subtitles.

The individual audio files were then programmatically combined based on their start and end times, with any gaps filled with silence. Special sound effects from the original audio track were manually transferred to the stitched audio track in Audacity. Finally, the synthesized audio track was added to the video file using VLC.

3.3 Subtitle Stacking

Video speech rates significantly impact comprehension, with faster dialogue often resulting in decreased understanding. Viewers fluent in the video's primary language tend to spend less time on subtitles [Szarkowska and Bogucka 2019], leaving those unfamiliar with the spoken language reliant on them for comprehension. Rapid speech can make it especially difficult for distracted viewers to keep up.

Subtitle stacking builds on conventional subtitle display systems that present captions sequentially at fixed times, extending them with persistence and accumulation to support recovery after missed content. To assist users, subtitles stack on the screen if previous ones are missed or still being read, continuing to layer up to a maximum of three lines before the video automatically pauses (Figure 2). Instead of using default timing, subtitles remain visible until the system detects that the user has finished reading, as indicated by gaze at the subtitle area. A viewer is considered distracted if they are not looking at the subtitle area.

The reading time for each subtitle was calculated using an average words-per-minute (WPM) measure. Subtitle reading time was estimated using an average speed of $r = 280$ WPM, with fixations computed per subtitle line as $T_{\text{read}} = \frac{N}{r/60}$. Each subtitle line was enclosed within a bounding box. Subtitles were removed once the viewer's gaze left the subtitle area and the cumulative fixation time reached T_{read}, where N is the number of words. This threshold is conservative for fast readers or viewers who skip subtitles [Muñoz 2017], allowing subtitles to be removed promptly for these viewers, while still giving slower readers sufficient time to read.

Table 1. Video stimuli presented in user study

Video Title	Primary Language	English Audio	Duration (min)
How LIGO discovered gravitational waves	Spanish	Synthesized	8:37
To understand autism, don't look away	Spanish	Synthesized	7:01
Our moral imperative to act on climate change	Italian	Original	8:41
Why journalists have an obligation to challenge power	Spanish	Synthesized	10:46

4 User Study

To compare the effectiveness of the intervention techniques when viewers are distracted, we conducted a comparative user study. Participants were tasked with watching four different videos using the gaze-aware video player with each of the intervention methods mentioned above, as well as one baseline (standard video player). While gaze-tracking methods could be misused for non-consensual attention monitoring, our system is designed for voluntary, transparent interaction. All collected data are anonymized, locally processed, and used solely for research purposes, as approved by the ethics committee.

4.1 Videos

We selected four TED talk videos, listed in Table 1. All the videos feature a single speaker presenting a story or educational talk. All the selected videos have a foreign language primary audio track. Where a secondary audio track was not present, an audio-dub was created in English as described in section 3.2. These videos were selected for their educational content, foreign language audio track, and minimal use of animations, which reduces the need to switch between visual content and subtitles, thereby minimizing cognitive load. This allows the viewer to focus on reading the subtitles without missing key information.

4.2 Participants

We recruited 24 participants through mass emails and targeted recruitment emails. Participants were from graduate and undergraduate levels with normal or corrected-to-normal vision. All participants were self-proclaimed fluent English speakers and were expected to have no difficulty with reading and writing, and none were fluent in Spanish or Italian. Of the 24 participants, 2 reported rarely using subtitles and 2 reported never using them when watching videos in a familiar language, while the remaining participants regularly used subtitles. All participants reported using their phones at least occasionally while watching videos, indicating that multitasking during video viewing was common. 17 participants self-reported having trouble keeping up with subtitles when watching videos. Sessions lasted approximately one hour, and all participants were compensated $20 for their participation.

4.3 Study Design

The user study was conducted in person at our institution. The study followed a within-subjects design with one primary independent variable, TECHNIQUE, with four levels (PAUSING, STACKING, DUBBING, and STANDARD). The STANDARD video player used the same interface, but with all gaze-based features disabled. The order of TECHNIQUE was counterbalanced across participants, while the order of videos remained fixed. TECHNIQUE was counterbalanced so that each participant encountered them in a different order, minimizing learning and order effects.

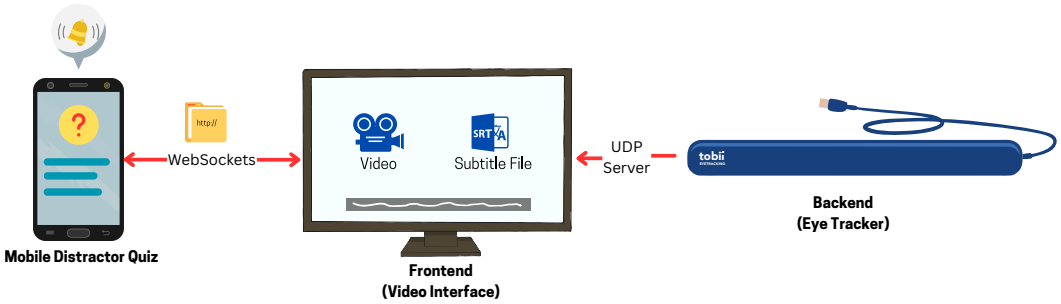

Fig. 3. System architecture including the mobile distractor quizzing component (left), the eye tracker listener (right), and the frontend video interface (middle).

The primary measures taken included the gaze interaction logs, distractor quiz scores, and comprehension quiz scores. Additionally, questionnaires provided subjective measures. Questionnaires and distractor quizzes are included in the supplementary material.

4.4 Apparatus and Software

The system comprises three main components: the frontend video player interface, the mobile quizzing interface, and the eye-tracker backend system. The three systems communicate through UDP and websockets (Figure 3). The source code is available at https://github.com/vialab/DontWannaMissAThing.

4.4.1 Hardware. A Tobii 4C eye tracker with a sampling rate of 90Hz was set up on a desktop equipped with 16GB RAM and a 24-inch external monitor featuring a 16:9 aspect ratio and a native resolution of 1920×1080 pixels. The eye tracker was calibrated once for each participant at the beginning of each session. Participants were seated with their heads approximately 60cm from the display, allowing them to move naturally and comfortably while maintaining accurate tracking. This provided participants with a sense of comfort and a more natural viewing experience.

4.4.2 Video Player Interface. The video player interface was developed in Electron to communicate with the eye-tracker. The video player layout is similar to a traditional video player with subtitles located at the bottom of the screen. Subtitle text was displayed in a white font colour on a semi-transparent (60%) black background. All three intervention methods used a similar interface, with the baseline (standard video player) using the same interface but with all gaze tracking features disabled. Subtitle text formatting and placement were kept consistent.

4.4.3 Backend. To communicate between the eye tracker and the video player interface, an application was developed using the Tobii SDK in C#. This backend application is used to subscribe to the gaze point data stream from the eye tracker and transfers the on-screen gaze point data as x-y pixel coordinates to the video player interface using UDP.

4.5 Procedure

Pre-Task. Participants were asked to complete a consent form and a pre-screening questionnaire. The pre-screening questions were used to ensure the participant was not familiar with the languages in the videos. Eye tracker calibration was performed for each participant before the user study began. The participant was then asked to connect to a mobile quizzing page on a provided mobile device and proceed to the first video.

Video Viewing. Each video was displayed using a standard video player or one of the three intervention methods. While watching the video stimuli, participants were routinely distracted by multiple-choice IQ questions [Carter 2008] on the provided mobile device. Distractor questions were presented every 30 to 60 seconds. Questions were chosen to require the participants' attention but could be completed quickly. Participants were instructed to answer the distractor questions as soon as they became aware of them. A notification sound was played as soon as the question appeared and would continue to ring if the question was not answered after 20 seconds. Each participant answered a total of 7 distractor questions per video. Immediately after completing the video, the participant was asked to complete a content comprehension quiz and evaluate the presented intervention method. This procedure was repeated until all three intervention methods had been tested, along with the baseline standard video player.

Post-Study. Finally, a post-study questionnaire was presented, and participants were asked to rate the four methods (three intervention methods and the standard player) based on a 5-point Likert scale.

5 Results

We report findings on the performance of the distractor and comprehension quizzes, followed by the participant perceptions of each intervention technique. Statistical analyses were conducted using the Friedman test and ART ANOVA [Wobbrock et al. 2011] due to normality violations, along with its post-hoc tests using the Holm-Bonferroni correction. We report 95% confidence intervals, estimated through bootstrapping with 10,000 iterations. All quizzes and questionnaires can be found in the supplementary material.

5.1 Distractor Quiz Performance

Participants were required to answer 7 multiple-choice IQ questions while watching the videos with each intervention technique. We measured their scores and the average time it took to answer the questions (Figure 4). Analysis showed a significant main effect of TECHNIQUE on the score ($F_{3,69} = 32.60$, $p < .001$, $\eta^2_G = .59$). Post-hoc test revealed that PAUSING significantly performed the best (M=98%, 95% CI:[97%, 100%]), followed by STANDARD (M=85%, 95% CI:[80%, 90%]), DUBBING (M=77%, 95% CI:[73%, 81%]), and STACKING (M=71%, 95% CI:[64%, 78%]). Further post-hoc test details can be found in Table 2. These results suggest that adaptive pausing effectively supports comprehension by mitigating the impact of distraction, whereas techniques involving audio or subtitle manipulation increase cognitive load and reduce accuracy.

We also measure the time between when the question was sent and when participants finished responding.TECHNIQUE did not significantly affect the time ($F_{3,69} = 2.63$, $p = .057$, $\eta^2_G = .10$). Participants responded in similar durations, with average completion times of 5.9s (95% CI:[5.6s, 6.1s])

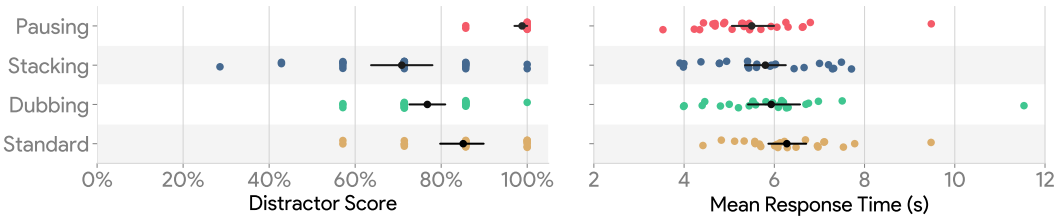

Fig. 4. Plots representing the distractor quiz performance across each intervention method. Black lines represent a 95% confidence interval. The plot suggests that PAUSING was the most effective method in mitigating distractions, resulting in the highest distractor scores. There was no significant difference in the mean time spent responding to the distractors.

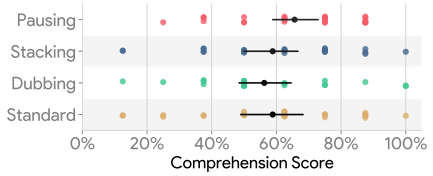

Fig. 5. Plot representing the comprehension scores across each intervention method. Black lines represent a 95% confidence interval. The plot suggests no significant difference in the scores.

across all conditions. Overall, participants completed the quizzes in comparable durations across all conditions, suggesting that the interventions did not impose additional time costs.

5.2 Comprehension Scores

Participants completed a comprehension test about each video after viewing it. We measured the scores for each participant (Figure 5). Analysis showed TECHNIQUE did not significantly affect the comprehension scores ($F_{3,69} = 1.05$, $p = .38$, $\eta^2_G = .04$). Participants had an average score of 60% (95% CI: [56%, 64%]) across the four quizzes. This suggests that while the interventions supported immediate task performance and attentional engagement during viewing, as reflected in the distractor quiz results, they did not lead to measurable differences in post-viewing comprehension. This suggests that their benefits may be limited to maintaining attention in real-time rather than enhancing long-term understanding.

5.3 Gaze Analysis

We analyzed eye-tracking data to assess how well each technique managed attention by measuring visual allocation and disengagement during distractions. Further details on the statistical test are in Appendix B.

Gaze Distribution. To quantify visual attention allocation, we analyzed the proportion of gaze points directed toward three mutually exclusive areas of interest (AOIs): *Subtitle*, *Video*, and *Distracted*. Analysis revealed a distinct preference for subtitles over visual content in foreign-language videos (Figure 6, left). We found no significant main effect of TECHNIQUE on gaze allocation ($F_{3,253} = 0.58$, $p = .628$, $\eta^2_G = .22$), indicating that the techniques did not alter the participants' attentional strategies. However, as expected, there was a significant main effect of AOI ($F_{2,253} = 241.04$, $p < .001$, $\eta^2_G = .96$). Subtitles attracted the highest proportion of gaze (M=59.7%, 95% CI: [56.0, 63.5]), followed by the video area (M=20.3%, 95% CI: [18.4, 22.3]) and distracted (M=12.6%, 95% CI: [10.3, 15.2]).

Gaze Shifts. We also quantified the frequency of *gaze shifts*, defined as moments when participants looked away from and subsequently returned to the video, during distractor questions (Figure 6, right). Analysis revealed significant main effects of TECHNIQUE on both the frequency of gaze shifts

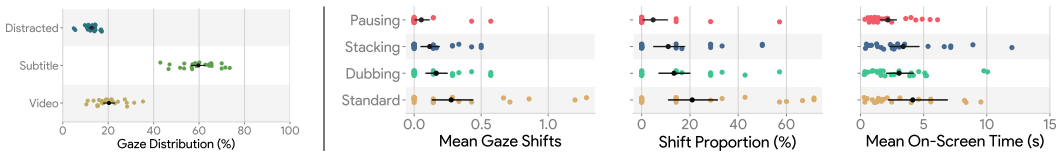

Fig. 6. Plots representing gaze behaviour during subtitled video viewing. The left plot shows the overall gaze distribution across three contexts: Video, Subtitle, and Distracted. The right three plots illustrate gaze behaviour during distractions for each technique. Plots suggest viewers prioritize subtitles, and PAUSING effectively reduces the attentional demands of standard playback, followed by STACKING and DUBBING. Black lines represent a 95% confidence interval.

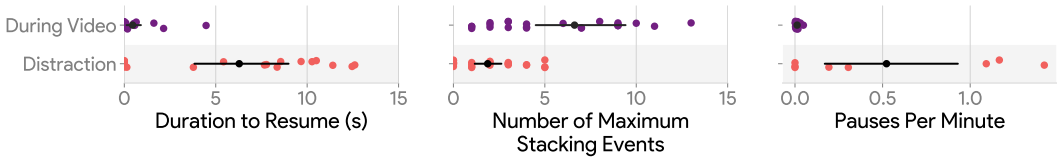

Fig. 7. Plots representing stacking behaviour with the STACKING technique. Black lines represent a 95% confidence interval. Stacking events occurred more frequently during video playback than during distraction periods, but pauses per minute remained low, suggesting that stacked subtitles generally fit within participants' natural reading speed and did not require pausing playback during distractions.

($F_{3,69} = 6.47$, $p < .001$, $\eta^2_G = .22$) and their likelihood ($F_{3,69} = 6.42$, $p < .001$, $\eta^2_G = .22$), measured as the proportion of distractor questions with gaze shifts. Post-hoc tests indicated that the PAUSING technique (M=5.4%, 95% CI : [0.6%, 11.9%]) significantly minimized gaze shifts compared to both STANDARD (M=27.6%, 95% CI : [11.2%, 31.6%]; $p < .001$) and DUBBING (M=16.4%, 95% CI : [8.5%, 24.7%]; $p < .01$). A significant main effect of TECHNIQUE ($F_{3,69} = 3.03$, $p < .05$, $\eta^2_G = .12$) on total on-screen duration during distractors was also observed. Participants spent significantly less time looking at the screen in the PAUSING condition (M=2.18s, 95% CI:[1.58s, 2.87s]) compared to STACKING (M=3.38s, 95% CI:[2.30s, 4.60s]; $p < .05$), though other differences were not significant. These findings suggest PAUSING was most effective in offloading visual attention. STACKING and DUBBING performed comparably as intermediate techniques, reducing visual demand relative to STANDARD. In contrast, STANDARD proved the most visually demanding, necessitating a persistent visual tether to the screen to avoid missing content.

Stacking Behaviour. We analyzed three metrics to understand STACKING behaviour: time to clear subtitles when paused, the number of maximum stacking events, and the frequency of video pauses (Figure 7). Participants required significantly longer ($F_{1,23} = 7.33$, $p < .05$, $\eta^2_G = .24$) to clear accumulated subtitles during distraction (M=6.28s, 95% CI:[3.85s, 9.00s]) compared to video-viewing (M=0.45s, 95% CI:[0.13s, 0.90s]). Regarding saturation events, we observed an inverse relationship between count and rate. The system reached its maximum capacity significantly more often ($F_{1,23} = 24.09$, $p < .001$, $\eta^2_G = .51$) in terms of total count during the video viewing phase. However, the rate of these events was significantly lower ($F_{1,23} = 20.29$, $p < .001$, $\eta^2_G = .47$) compared to the distraction phase. This discrepancy is attributed to the significantly longer duration of the video viewing phase compared to the distraction periods. As expected, prolonged visual disengagement during distractor questions led to substantial subtitle accumulation, resulting in longer recovery times. In contrast, video viewing resulted in a higher total number of maximum events, but they were momentary, suggesting that the reading threshold frequently aligned with participants' natural reading speed.

5.4 Subjective Assessment for Attention and Recovery

Participants rated the effectiveness of the intervention techniques in supporting attention and recovery from distraction on a 5-point Likert scale (Figure 8). Further details on the statistical test can be found in Appendix C.

Attention Recovery. Participants rated how effectively each intervention technique helped them recover from distractions. The Friedman test indicated a significant effect of TECHNIQUE on perceived attention recovery ($\chi^2_F(3) = 34.80$, $p < .0001$). Post-hoc Wilcoxon tests showed that all three gaze-aware interventions were rated significantly higher than STANDARD ($p < .001$). PAUSING and STACKING received the highest ratings (MDN=4.0). No significant difference was found between PAUSING and

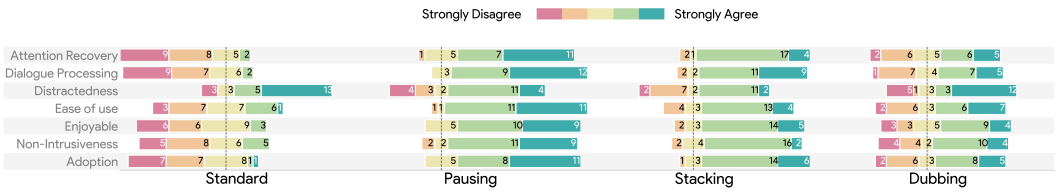

Fig. 8. Questionnaire responses for each intervention method. Participants thought pausing and stacking were the most effective in mitigating distractions, followed by dubbing. Standard performed the worst, suggesting that gaze-aware interventions support attention recovery and a more seamless viewing experience.

STACKING ($p = .334, n.s.$), suggesting comparable perceived support for regaining focus. However, DUBBING (MDN=3.0) was rated significantly lower than the other two intervention methods ($p < .05$).

Dialogue Processing. The extent to which participants felt able to keep up with the video's dialogue. A significant main effect of TECHNIQUE was found ($\chi^2_F(3) = 37.80, p < .001$), where PAUSING (MDN=4.5), STACKING (MDN=4.0), and DUBBING (MDN=3.5) rated higher than STANDARD (MDN=2.0). PAUSING and STACKING were also rated significantly better than DUBBING ($p < .05$), indicating that adaptive pausing and subtitle stacking were perceived as most helpful for maintaining comprehension.

Distractedness. Degree to which participants felt distracted from the video while answering distractor questions. A significant main effect of TECHNIQUE was found ($\chi^2_F(3) = 13.82, p < .05$). Post-hoc tests showed that STACKING (MDN=4.0) significantly reduced perceived distraction compared to the STANDARD (MDN=5.0). No other pairwise comparisons reached significance ($p = .203, n.s.$). This suggests that stacking subtitles was perceived as the most effective intervention for reducing distraction, although the overall differences were moderate.

User Experience. Perceived enjoyment and how easy or difficult participants found the technique to use. Enjoyment ratings differed significantly across TECHNIQUE ($\chi^2_F(3) = 25.91, p < .001$). Both PAUSING and STACKING were rated more enjoyable than STANDARD (MDN=5.0 vs. 2.5). PAUSING was also rated significantly easier to use (MDN=4.0) compared to STANDARD (MDN=3.0) ($\chi^2_F(3) = 21.45, p < .001$). These results suggest that adaptive interventions not only supported engagement but also maintained usability, contributing to a more positive overall viewing experience.

User Acceptance. Degree to which the technique was perceived as distracting, and willingness to use the technique if available. Perceived distraction (non-intrusiveness) showed a significant main effect ($\chi^2_F(3) = 23.15$, $p < .001$). PAUSING and STACKING were perceived as less distracting than STANDARD (MDN=4.0 vs. 2.0). Finally, willingness to use (adoption) also differed significantly ($\chi^2_F(3) = 29.77, p < .001$). All three adaptive methods were preferred over STANDARD (MDN=4.0, $p < .05$), with PAUSING and STACKING (MDN=4.0) again leading. These findings indicate that gaze-adaptive methods, which minimize disruption to the viewing flow, are more likely to be accepted by users in real-world scenarios.

Summary. Overall, participants consistently rated PAUSING and STACKING as providing stronger attentional support, higher enjoyment, and better usability than the STANDARD player. DUBBING was generally rated above STANDARD but below the other intervention methods. These results suggest that adaptive pausing and stacked subtitles effectively supported attention recovery and sustained comprehension, aligning with participants' subjective perceptions of engagement and control.

5.5 Attention Recovery and Viewing Experience

We report three qualitative themes that capture how the techniques affected attention recovery, comprehension, and the overall viewing experience, as expressed in participants' comments from open-ended questions in post-task questionnaires.

Supporting Focus and Reducing Missed Content. A common theme was that the interventions helped participants stay aligned with the video and avoid missing content. Many appreciated that the techniques worked for

them when their attention lapsed. Participants described pausing as "really helpful"[P8] and "the technique that helped me the most"[P16], noting that "the video stopped when I wasn't looking, so that was great" [P7]. Others appreciated the reassurance that they were not missing key material ("I had the audio feedback that I didn't miss anything from the video" [P11]).

Stacked subtitles similarly provided reassurance, allowing viewers to catch up after a distraction. Participants said it was "easy to catch up when looking back" [P4] and that keeping previous lines visible "ensured that I was reading all the information from the video" [P16]. This adaptation made it possible to reconstruct what was missed without having to rewind the video. Even with the dubbing technique, some noted that hearing an alternate-language track while distracted "helped me stay on track" [P20] and "made it possible for me to understand what she was saying" [P14].

In contrast, the baseline underscored the importance of these adaptive features: without any intervention, participants reported feeling "confused" [P6], "not really focused" [P14], and that they "lost the flow every single time" [P21]. These responses underscore the role of gaze-awareness in providing participants with a sense of continuity and control during divided attention.

Responsiveness and Natural Flow. While participants appreciated interventions that quickly addressed distractions, they also noted some nuances that can disrupt the viewing experience. Some found gaze-triggered pauses too abrupt or overly sensitive ("It can be jarring when the video pauses when you aren't "fully" distracted" [P18]). Others desired a short buffer or replay window ("start playing from ten seconds before stop" [P5]) to create smoother transitions. This tension also appeared in the stacked subtitle condition, where participants experienced inconsistencies in timing and gaze detection. A few felt "the video took too long to pause" [P11], while others said "the subtitles were too slow for my reading speed" [P23].

With dubbing, participants pointed out that "switching from reading to listening was challenging" [P18]. Others commented that transitions felt "a bit off putting" [P16] or "slightly annoying instead of being helpful" [P2]. At the same time, some participants found the technique helpful, noting that "providing an audio recording in English when I was answering questions on the mobile phone helped me stay on track" [P20] and that the system was "really effective" [P7]. Across techniques, participants sought adaptive responses that felt seamless yet not disruptive to the viewing experience.

Cognitive Load and Attention Shifts. The interventions also revealed how modality and task-switching affect cognitive effort. Participants described the pausing technique as "sorting out tasks very easily" [P21], reducing the stress of multitasking by explicitly segmenting attention between video and phone. Stacked subtitles also reduced the stress, where one noted "occasionally when looking away, a single caption would be missed, but this provided the easiest way to keep up with the content" [P17].

Dubbing, which required simultaneous listening and answering, drew the strongest reactions around mental load. Participants frequently reported that "[my] mind was trying to concentrate on two things" [P5] and that "changing voice and language was distracting" [P3]. The comments suggest that the challenge arises more from the effort required to switch between modalities. Gaze-aware systems may unintentionally increase cognitive demands when transitions between reading and listening are misaligned with users' natural processing strategies.

6 Discussion

Our results show that gaze-adaptive playback can effectively support attention recovery during subtitle-based video viewing. The discussion interprets these findings in relation to real-time adaptation, modality, and individual viewing behaviour, highlighting how gaze-driven interventions shape immediate recovery, cognitive effort, and user experience.

Gaze-adaptive interventions support real-time recovery, not long-term comprehension. The study was intentionally conducted under controlled conditions to isolate the perceptual and attentional effects of gaze-triggered interventions before extending to in-the-wild viewing scenarios. The findings of this study demonstrate that gaze-adaptive interventions effectively help viewers recover from distractions that lead to immediate lapses in attention during subtitle reading. These gaze-adaptive interventions, such as adaptive pausing and persistent subtitles, reduce missed information by maintaining or reintroducing missed content when a distraction is detected. Our gaze shift analysis shows that the PAUSING technique reduced off-screen

checks to 5.4%, compared to 27.6% in the STANDARD condition. By decoupling playback from time, users could disengage during distractions without the cognitive burden of monitoring the video. Recovering from distractions addresses the limitations of subtitles, as videos do not allow for selective re-reading like text does. Ultimately, this benefits viewers by offering an alternative to relying on explicit recovery strategies, such as scrubbing the timeline. While our gaze analysis captures attention dynamics during distractions, it does not directly characterize post-distraction recovery behaviour (*e.g.*, re-fixation latency on subtitles or subtitle re-reading), which remains for future work.

The short-term benefits of real-time attention recovery do not translate into long-term comprehension benefits. The benefits of having access to recent information include supporting immediate task performance and reducing cognitive load, but they do not support the development of comprehension on their own. This is likely due to the nature of the stimuli used in our study, which consisted of narrative and informational TED-style talks rather than instructional materials that build cumulative understanding. As such, the interventions primarily helped viewers stay temporally synchronized with the content but did not reinforce conceptual learning or memory. This contrasts with findings from educational contexts, where gaze-aware interventions are structured to promote learning and deeper comprehension [Mills et al. 2021; Santhosh et al. 2024].

Modality changes support recovery only when matched to the viewer's preference. The adaptation methods emphasize how modality and cognitive load influence attention recovery. Prior work on bilingual audiovisual learning shows that audio presence, subtitle language, and language proficiency influence attentional allocation and lexical processing [Abu-Rayyash et al. 2024]. This suggests that abrupt modality shifts can overload working memory when viewers manage dual-language processing. Strategies that reduce the simultaneous mental processing of audio and visual, such as pausing the video to keep missed subtitles visible, often improve task performance and are less disruptive for viewers. However, the presentation format is important. While STACKING maintained visual context, participants spent more time looking at the screen compared to the PAUSING condition. This indicates that a continuous video flow, even with subtitles, keeps users engaged, preventing the attentional offloading seen with strict pauses. In contrast, approaches that shift modality quickly can require additional mental effort, particularly for non-fluent viewers who rely on subtitles to help them comprehend the content.

The results demonstrate that gaze adaptive recovery is effective, based on the timing of the system's intervention and the presentation of information to the viewer. To reduce cognitive load during recovery, adaptive playback systems should provide viewers with sufficient time to process missed material, rather than presenting denser information. Modality-based adaptations, such as switching modalities to audio, may support specific users. Our data on DUBBING reflects this nuance: while it reduced gaze shifts relative to the STANDARD condition, it did not eliminate checking behaviour entirely, suggesting that users still felt a need to visually verify context despite the auditory support. The effectiveness relies on balancing sensory diversity and cognitive load.

Attention-aware interventions must align with personal viewing behaviour. To achieve a seamless gaze-adaptive playback experience, the gaze-adaptive system must mitigate the tension between sensitivity and stability. If the system does not intervene to minimize disruptions at a smooth and predictable rate of change, the viewer will find it to be disruptive. Viewers prefer smooth, predictable, and minimally intrusive interventions. Intervention techniques, such as implementing buffered pause, gradual transitions, and replay windows, can maintain and flow through the video while allowing time for the individual to recover.

Personalization is also essential as individuals vary in their reading speeds, attention patterns, and viewing behaviours. Due to this variability, it is insufficient to establish fixed thresholds for gaze-related responses. The stacking analysis showed that the system sometimes reached maximum capacity during normal viewing rather than during distractions. This indicates a misalignment between fixed thresholds and natural reading speeds, leading to unnecessary disruptions in the viewing flow. Lightweight calibration and adjustable features, such as pause perception and subtitle duration, give the system more flexibility in combination with the viewer's viewing behaviour. The effectiveness of the gaze-adaptive interface stems from a combination of accurate gaze detection and the system's conscious respect for the individual's level of attention.

7 Limitations & Future Work

Several limitations restrict the generalizability of our results. First, the order of techniques was counterbalanced, but the videos were shown in a fixed sequence, which may leave residual content or ordinal effects. Also, this study took place in a controlled environment using short, subtitled video-based interventions and a fixed-gaze-based attention model. Although this setup allowed comparisons of interventions, it may not fully reflect natural viewing contexts, where distraction types, durations, and frequencies vary. Future research should investigate how gaze-adaptive recovery translates to other contexts, such as extended video, mobile devices, or multitasking environments.

Secondly, all interventions in this study were designed and tested for a single viewer. While this allowed us to systematically evaluate gaze-adaptive recovery in a controlled setting, it does not account for scenarios with multiple viewers. In multi-viewer situations, interventions could be extended in several ways. For example, while playback continues, stacked subtitles can be displayed, and once a limit is reached, a transition can occur to an AI-generated summary for viewers who have missed the content. Alternatively, personalized catch-up notifications could be sent to a phone, allowing each viewer to recover missed information without disrupting others' viewing experience. Similarly, subtitles could be transferred directly to a mobile device, enabling viewers to follow along independently or asynchronously. Future work should explore how gaze-adaptive interventions can be scaled to multi-viewer contexts and mobile-assisted recovery while maintaining a seamless viewing experience.

Thirdly, our participant sample primarily consisted of viewers who were non-fluent in the video language (Spanish or Italian) but fluent in English. However, language proficiency, familiarity with subtitled content, and viewing behaviours all affect attention and recovery. Future work should consider more diverse participant samples and languages (subtitles and video) to determine how gaze-adaptive systems generalize across different linguistic contexts.

Lastly, while our interventions were effective in supporting short-term recovery, they were not intended to promote comprehension or retention over longer periods, given the nature of the video stimuli. Future studies should consider hybrid systems that combine real-time adaptation and educational support to facilitate immediate attention recovery in educational contexts. Furthermore, the attention-based dubbing intervention used time-stretching synthesized speech to align with subtitle timings, which are often condensed or expanded compared to spoken dialogue. This could have led to slower or faster-than-normal speech rates, affecting users' perceptions of the dubbed audio.

8 Conclusion

Distractions often disrupt subtitle-based viewing, making it difficult for non-fluent viewers to recover. We explored three gaze-aware interventions to help viewers recover from distraction: adaptive pausing, stacked subtitles, and dubbing. A user study found adaptive pausing to be the most effective, as it improved task performance and reduced distractions. Stacked subtitles aided recovery but may slow fast readers, while dubbing provided limited support and increased cognitive load for some. All interventions outperformed standard playback during immediate recovery but did not enhance long-term comprehension. These results highlight the potential of gaze-adaptive methods and the importance of user engagement and preferences. This work lays the foundation for gaze-adaptive systems that seamlessly integrate attention recovery into everyday viewing experiences.

Acknowledgments

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). We also thank the participants in this study.

References

Hussein Abu-Rayyash, Shatha Alhawamdeh, and Yuri Ringomon. 2024. The eye-ear relationship: investigating auditory impacts on subtitle reading and comprehension. *Texto Livre* 17 (2024), e52687. doi:10.1590/1983-3652.2024.52687

Emad A. Alghamdi, Paul Gruba, and Eduardo Velloso. 2022. The Relative Contribution of Language Complexity to Second Language Video Lectures Difficulty Assessment. *The Modern Language Journal* 106, 2 (2022), 393–410. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/modl.12773 doi:10.1111/modl.12773

Riku Arakawa and Hiromu Yakura. 2021. Mindless Attractor: A False-Positive Resistant Intervention for Drawing Attention Using Auditory Perturbation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 99, 15 pages. doi:10.1145/3411764.3445339

Ryan S.J.d. Baker, Sidney K. D'Mello, Ma.Mercedes T. Rodrigo, and Arthur C. Graesser. 2010. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies* 68, 4 (2010), 223–241. doi:10.1016/j.ijhcs.2009.12.003

Karolina Baranowska. 2020. Learning most with least effort: subtitles and cognitive load. *ELT Journal* 74, 2 (03 2020), 105–115. doi:10.1093/elt/ccz060

Jonathan Bidwell and Henry Fuchs. 2011. Classroom analytics: Measuring student engagement with automated gaze tracking. *Behav Res Methods* 49, 113 (2011), 17 pages. doi:10.13140/RG.2.1.4865.6242

Giulia Borghini and Valerie Hazan. 2018. Listening Effort During Sentence Processing Is Increased for Non-native Listeners: A Pupillometry Study. *Frontiers in Neuroscience* Volume 12 - 2018 (2018), 13 pages. doi:10.3389/fnins.2018.00152

Diane M. Bunce, Elizabeth A. Flens, and Kelly Y. Neiles. 2010. How Long Can Students Pay Attention in Class? A Study of Student Attention Decline Using Clickers. *Journal of Chemical Education* 87, 12 (2010), 1438–1443. arXiv:https://doi.org/10.1021/ed100409p doi:10.1021/ed100409p

Philip J Carter. 2008. *Advanced IQ Tests: The Toughest Practice Questions to Test Your Lateral Thinking, Problem Solving and Reasoning Skills.* Kogan Page Limited, London, UK.

Wing Shan Chan, Jan-Louis Kruger, and Stephen Doherty. 2022. An investigation of subtitles as learning support in university education. *The Journal of Specialised Translation* 38, 38 (Jul. 2022), 155–179. doi:10.26034/cm.jostrans.2022.087

Brendan David-John, Candace Peacock, Ting Zhang, T. Scott Murdison, Hrvoje Benko, and Tanya R. Jonker. 2021. Towards gaze-based prediction of the intent to interact in virtual reality. In *ACM Symposium on Eye Tracking Research and Applications* (Virtual Event, Germany) *(ETRA '21 Short Papers)*. Association for Computing Machinery, New York, NY, USA, Article 2, 7 pages. doi:10.1145/3448018.3458008

Sidney D'Mello, Andrew Olney, Claire Williams, and Patrick Hays. 2012. Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human-Computer Studies* 70, 5 (2012), 377–398. doi:10.1016/j.ijhcs.2012.01.004

Andrew Duchowski, Patrícia Vieira, Ítalo Assis, Krzysztof Krejtz, Chris Hughes, and Pilar Orero. 2025. Interactive Storytelling with Gaze-Responsive Subtitles. In *Proceedings of the ACM International Conference on Interactive Media Experiences Workshops* (Niterói/RJ). SBC, Porto Alegre, RS, Brasil, 19–25. doi:10.5753/imxw.2025.9779

Géry d'Ydewalle and Wim De Bruycker. 2007. Eye movements of children and adults while reading television subtitles. *European psychologist* 12, 3 (2007), 196–205.

Marco Porta Elisa Perego, Fabio Del Missier and Mauro Mosconi. 2010. The Cognitive Effectiveness of Subtitle Processing. *Media Psychology* 13, 3 (2010), 243–272. doi:10.1080/15213269.2010.502873

Stephen Hutt, Caitlin Mills, Nigel Bosch, Kristina Krasich, James Brockmole, and Sidney D'Mello. 2017. "Out of the Fr-Eye-ing Pan": Towards Gaze-Based Models of Attention during Learning with Technology in the Classroom. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization* (Bratislava, Slovakia) *(UMAP '17)*. Association for Computing Machinery, New York, NY, USA, 94–103. doi:10.1145/3079628.3079669

Aulikki Hyrskykari, Päivi Majaranta, and Kari-Jouko Räihä. 2005. From Gaze Control to Attentive Interfaces. In *Proceedings of HCII 2005*. Lawrence Erlbaum Associates, Mahwah, NJ, USA, 10 pages.

Jan-Louis Kruger, Esté Hefer, and Gordon Matthew. 2013. Measuring the impact of subtitles on cognitive load: eye tracking and dynamic audiovisual texts. In *Proceedings of the 2013 Conference on Eye Tracking South Africa* (Cape Town, South Africa) *(ETSA '13)*. Association for Computing Machinery, New York, NY, USA, 62–66. doi:10.1145/2509315.2509331

Jan-Louis Kruger and Faans Steyn. 2013. Subtitles and Eye Tracking: Reading and Performance. *Reading Research Quarterly* 49 (10 2013). doi:10.1002/rrq.59

Kuno Kurzhals, Michael Burch, Tanja Blascheck, Gennady Andrienko, Natalia Andrienko, and Daniel Weiskopf. 2017. A Task-Based View on the Visual Analysis of Eye-Tracking Data. In *Eye Tracking and Visualization*, Michael Burch, Lewis Chuang, Brian Fisher, Albrecht Schmidt, and Daniel Weiskopf (Eds.). Springer International Publishing, Cham, 3–22.

Kuno Kurzhals, Fabian Göbel, Katrin Angerbauer, Michael Sedlmair, and Martin Raubal. 2020. A View on the Viewer: Gaze-Adaptive Captions for Videos. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3313831.3376266

Seungyeon Lee, Ian M. McDonough, Jessica S. Mendoza, Mikenzi B. Brasfield, Tasnuva Enam, Catherine Reynolds, and Benjamin C. Pody. 2021. Cellphone addiction explains how cellphones impair learning for lecture materials. *Applied Cognitive Psychology* 35, 1 (2021), 123–135. doi:10.1002/acp.3745

Sixin Liao, Lili Yu, Jan-Louis Kruger, and Erik D. Reichle. 2022. The impact of audio on the reading of intralingual versus interlingual subtitles: Evidence from eye movements. *Applied Psycholinguistics* 43, 1 (2022), 237–269. doi:10.1017/S0142716421000527

Sophie I. Lindquist and John P. McLean. 2011. Daydreaming and its correlates in an educational environment. *Learning and Individual Differences* 21, 2 (2011), 158–167. doi:10.1016/j.lindif.2010.12.006

Paul. L. Markham, Lizette A. Peter, and Teresa J. McCarthy. 2001. The Effects of Native Language vs. Target Language Captions on Foreign Language Students' DVD Video Comprehension. *Foreign Language Annals* 34, 5 (2001), 439–445. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1944-9720.2001.tb02083.x doi:10.1111/j.1944-9720.2001.tb02083.x

Jacek Matulewski, Bibianna Bałaj, Ewelina Marek, Łukasz Piasecki, Dawid Gruszczyński, Mateusz Kuchta, and Włodzisław Duch. 2018. Moveye: gaze control of video playback. In *Proceedings of the Workshop on Communication by Gaze Interaction* (Warsaw, Poland) *(COGAIN '18)*. Association for Computing Machinery, New York, NY, USA, Article 4, 5 pages. doi:10.1145/3206343.3206352

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. doi:10.1080/07370024.2020.1716762

Holger Mitterer and James M. McQueen. 2009. Foreign Subtitles Help but Native-Language Subtitles Harm Foreign Speech Perception. *PLOS ONE* 4, 11 (11 2009), 1–5. doi:10.1371/journal.pone.0007785

Carmen Muñoz. 2017. The role of age and proficiency in subtitle reading. An eye-tracking study. *System* 67 (2017), 77–86. doi:10.1016/j.system.2017.04.015

Shivsevak Negi and Ritayan Mitra. 2020. Fixation Duration and the Learning Process: An Eye Tracking Study with Subtitled Videos. *Journal of Eye Movement Research* 13, 6 (2020), 1–15. doi:10.16910/jemr.13.6.1

Cuong Nguyen and Feng Liu. 2016. Gaze-based Notetaking for Learning from Lecture Videos. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2093–2097. doi:10.1145/2858036.2858137

Amy Pavel, Colorado Reed, Björn Hartmann, and Maneesh Agrawala. 2014. Video digests: a browsable, skimmable format for informational lecture videos. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, Hawaii, USA) *(UIST '14)*. Association for Computing Machinery, New York, NY, USA, 573–582. doi:10.1145/2642918.2647400

Z. Ellen Peng and Lily M. Wang. 2016. Effects of noise, reverberation and foreign accent on native and non-native listeners' performance of English speech comprehension. *The Journal of the Acoustical Society of America* 139, 5 (05 2016), 2772–2783. doi:10.1121/1.4948564

Elisa Perego, Fabio Del Missier, Marco Porta, and Mauro Mosconi. 2010. The Cognitive Effectiveness of Subtitle Processing. *Media Psychology* 13, 3 (2010), 243–272. arXiv:https://doi.org/10.1080/15213269.2010.502873 doi:10.1080/15213269.2010.502873

Martin J. Pickering, Steven Frisson, Brian McElree, and Matthew J. Traxler. 2004. Eye Movements and Semantic Composition. In *On-line Study of Sentence Comprehension: Eyetracking, ERPs and Beyond*, Manuel Carreiras and Charles Jr. Clifton (Eds.). Psychology Press, New York, NY, USA, 33–50.

Jayasankar Santhosh, Andreas Dengel, and Shoya Ishimaru. 2024. Gaze-Driven Adaptive Learning System With ChatGPT-Generated Summaries. *IEEE Access* 12 (2024), 173714–173733. doi:10.1109/ACCESS.2024.3503059

Kshitij Sharma, Hamed S. Alavi, Patrick Jermann, and Pierre Dillenbourg. 2016. A gaze-based learning analytics model: in-video visual feedback to improve learner's attention in MOOCs. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (Edinburgh, United Kingdom) *(LAK '16)*. Association for Computing Machinery, New York, NY, USA, 417–421. doi:10.1145/2883851.2883902

Agnieszka Szarkowska and Lidia Bogucka. 2019. Six-second rule revisited. *Translation, Cognition & Behavior* 2, 1 (2019), 101–124. doi:10.1075/tcb.00022.sza

Chinchu Thomas and Dinesh Babu Jayagopi. 2017. Predicting student engagement in classrooms using facial behavioral cues. In *Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education* (Glasgow, UK) *(MIE 2017)*. Association for Computing Machinery, New York, NY, USA, 33–40. doi:10.1145/3139513.3139514

Narayanan Veliyath, Pradipta De, Andrew A. Allen, Charles B. Hodges, and Aniruddha Mitra. 2019. Modeling Students' Attention in the Classroom using Eyetrackers. In *Proceedings of the 2019 ACM Southeast Conference* (Kennesaw, GA, USA) *(ACMSE '19)*. Association for Computing Machinery, New York, NY, USA, 2–9. doi:10.1145/3299815.3314424

Roel Vertegaal. 2002. Designing attentive interfaces. In *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications* (New Orleans, Louisiana) *(ETRA '02)*. Association for Computing Machinery, New York, NY, USA, 23–30. doi:10.1145/507072.507077

Nigel G. Ward, Chelsey N. Jurado, Ricardo A. Garcia, and Florencia A. Ramos. 2016. On the possibility of predicting gaze aversion to improve video-chat efficiency. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications* (Charleston, South Carolina) *(ETRA '16)*. Association for Computing Machinery, New York, NY, USA, 267–270. doi:10.1145/2857491.2857497

Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems.*

ACM, New York, NY, USA, 143–146. doi:10.1145/1978942.1978963

Xiang Xiao and Jingtao Wang. 2016. Context and cognitive state triggered interventions for mobile MOOC learning. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (Tokyo, Japan) *(ICMI '16)*. Association for Computing Machinery, New York, NY, USA, 378–385. doi:10.1145/2993148.2993177

Xiang Xiao and Jingtao Wang. 2017. Understanding and Detecting Divided Attention in Mobile MOOC Learning. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2411–2415. doi:10.1145/3025453.3025552

Janez Zaletelj and Andrej Košir. 2017. Predicting students' attention in the classroom from Kinect facial and body features. *EURASIP Journal on Image and Video Processing* 2017, 1 (2017), 80. doi:10.1186/s13640-017-0228-8

Yueyuan Zheng, Xinchen Ye, and Janet Hsiao. 2019. Does Video Content Facilitate or Impair Comprehension of Documentaries? The Effect of Cognitive Abilities and Eye Movement Strategy. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*. The Cognitive Science Society, Austin, TX, USA, 1231–1237.

A Statistical Analysis on Distractor Scores

Table 2. ART ANOVA analysis of the main effects and post-hoc tests for distractor scores.

DISTRACTOR ($F_{3,69} = 32.60$, $p < .001$, $\eta^2_G = 0.59$)				
comparisons		*Mean diff (%)*	*p*	
STANDARD	PAUSING	-13.7	< .001	***
STANDARD	STACKING	14.3	< .001	***
STANDARD	DUBBING	8.3	< .01	**
PAUSING	STACKING	28.0	< .001	***
PAUSING	DUBBING	22.0	< .001	***
STACKING	DUBBING	-6.0	.363	

B Statistical Analysis on Gaze Analysis

Table 3. Comparisons for the number of gaze shifts.

GAZE SHIFTS ($F_{3,69} = 6.47$, $p < .001$, $\eta^2_G = .22$)				
comparisons		*Mean diff*	*p*	
STANDARD	PAUSING	0.22	< .001	***
STANDARD	STACKING	0.16	.064	
STANDARD	DUBBING	0.11	.433	
PAUSING	STACKING	-0.06	.064	
PAUSING	DUBBING	-0.11	< .01	**
STACKING	DUBBING	-0.05	.249	

Table 4. Comparisons for the gaze shift likelihood.

SHIFT PROPORTION ($F_{3,69} = 6.42$, $p < .001$, $\eta^2_G = .22$)				
comparisons		*Mean diff (%)*	*p*	
STANDARD	PAUSING	16.2	< .001	***
STANDARD	STACKING	16.0	.077	
STANDARD	DUBBING	11.2	.354	
PAUSING	STACKING	-6.2	.064	
PAUSING	DUBBING	-11.0	< .01	**
STACKING	DUBBING	-4.8	.354	

Table 5. Comparisons for on-screen time during distractions.

ON-SCREEN TIME ($F_{3,69} = 3.03$, $p < .05$, $\eta^2_G = .12$)				
comparisons		*Mean diff (s)*	*p*	
STANDARD	PAUSING	1.96	.111	
STANDARD	STACKING	0.75	.533	
STANDARD	DUBBING	1.09	.842	
PAUSING	STACKING	-1.21	< .05	*
PAUSING	DUBBING	-0.87	.116	
STACKING	DUBBING	0.34	.504	

C Statistical Analysis on Perceived Support for Attention and Recovery

Table 6. Comparisons for attention recovery.

ATTENTION RECOVERY ($\chi^2_F(3) = 34.80, p < .001$)				
comparisons		*Mean diff*	*p*	
STANDARD	PAUSING	-2.17	< .001	***
STANDARD	STACKING	-1.96	< .001	***
STANDARD	DUBBING	-1.25	< .01	**
PAUSING	STACKING	0.21	.334	
PAUSING	DUBBING	0.92	< .05	*
STACKING	DUBBING	0.71	< .05	*

Table 7. Comparisons for dialogue processing.

DIALOGUE PROCESSING ($\chi^2_F(3) = 37.80, p < .001$)				
comparisons		*Mean diff*	*p*	
STANDARD	PAUSING	-2.33	< .001	***
STANDARD	STACKING	-2.08	< .001	***
STANDARD	DUBBING	-1.29	< .01	**
PAUSING	STACKING	0.25	.330	
PAUSING	DUBBING	1.04	< .01	**
STACKING	DUBBING	0.79	< .05	*

Table 8. Comparisons for distractedness.

DISTRACTEDNESS ($\chi^2_F(3) = 13.82, p < .05$)				
comparisons		*Mean diff*	*p*	
STANDARD	PAUSING	0.71	.203	
STANDARD	STACKING	0.88	< .05	*
STANDARD	DUBBING	0.38	.923	
PAUSING	STACKING	0.17	.923	
PAUSING	DUBBING	-0.33	.923	
STACKING	DUBBING	-0.50	.259	

Table 9. Comparisons for ease of use.

EASE OF USE ($\chi^2_F(3) = 21.45, p < .001$)				
comparisons		*Mean diff*	*p*	
STANDARD	PAUSING	-1.54	< .01	**
STANDARD	STACKING	-0.92	< .05	*
STANDARD	DUBBING	-0.63	.137	
PAUSING	STACKING	0.63	.062	
PAUSING	DUBBING	0.92	< .05	*
STACKING	DUBBING	0.29	.331	

Table 10. Comparisons for enjoyment.

ENJOYMENT ($\chi^2_F(3) = 25.91, p < .001$)				
comparisons		*Mean diff*	*p*	
STANDARD	PAUSING	-1.79	< .001	***
STANDARD	STACKING	-1.54	< .001	***
STANDARD	DUBBING	-0.96	< .05	*
PAUSING	STACKING	0.25	.311	
PAUSING	DUBBING	0.83	< .05	*
STACKING	DUBBING	0.58	.096	

Table 11. Comparisons for non-intrusiveness.

NON-INTRUSIVENESS ($\chi^2_F(3) = 23.15, p < .001$)				
comparisons		*Mean diff*	*p*	
STANDARD	PAUSING	-1.67	< .01	**
STANDARD	STACKING	-1.29	< .01	**
STANDARD	DUBBING	-0.79	.138	
PAUSING	STACKING	0.38	.138	
PAUSING	DUBBING	0.88	.059	
STACKING	DUBBING	0.50	.138	

Table 12. Comparisons for adoption.

ADOPTION ($\chi^2_F(3) = 29.77, p < .001$)				
comparisons		*Mean diff*	*p*	
STANDARD	PAUSING	-2.00	< .001	***
STANDARD	STACKING	-1.79	< .001	***
STANDARD	DUBBING	-1.08	< .05	*
PAUSING	STACKING	0.21	.398	
PAUSING	DUBBING	0.92	< .05	*
STACKING	DUBBING	0.71	< .05	*