Automated Detection of Attention and Retention in Educational Videos Using Eye-Tracking, Dynamic Areas of Interest and Feature Fusion

Sébastien Lallé, Sina Nikneshan, Solène Lambert, Vanda Luengo, Ali Abou-Hassan

Automated Detection of Attention and Retention in Educational Videos Using Eye-Tracking, Dynamic Areas of Interest and Feature Fusion

Sébastien Lallé[1], Sina Nikneshan[1,2], Solène Lambert[3,4], Vanda Luengo[1], and Ali Abou-Hassan[5,6]

[1]Sorbonne University, CNRS, LIP6, F-75005 Paris, France
[2]Université Paris-Est Créteil, 94000 Créteil, France
[3]Sorbonne University, INSERM, ISIR, F-75005 Paris, France
[4]Sorbonne University, CAPSULE, F-75005 Paris, France
[5]Sorbonne University, CNRS, PHENIX, F-75005 Paris, France
[6]Institut Universitaire de France (IUF), 75231 Paris, France
{sebastien.lalle, sina.nikneshan, solene.lambert, vanda.luengo, ali.abou_hassan}@sorbonne-universite.fr

Abstract. Educational videos are widely used in remote and blended learning. However, learners' attention often fluctuates while watching, which can hinder their retention of key information. This, in turn, may impact their overall learning outcomes. Detecting when learners lose attention or fail to memorize key elements of a video could help address these challenges—for example, by enabling adaptive support that enhances engagement and retention, bridging the gap between passive video consumption and active learning. Such automated detection could also provide valuable insights to instructors about when attention and retention drop in their videos. In this study, we explore how to detect learners' attention and retention while they watch an educational video in a blended course on green chemistry, using eye-tracking data. To achieve this, we develop machine learning classifiers that analyze eye movements, pupil dilation, eye-screen distance, and attention to dynamically tracked areas of interest (AOIs). We investigate different strategies for fusing these types of information and find that dynamic AOIs can significantly improve ML predictions, albeit with moderate performance.

Keywords: Educational video · Eye tracking · Classification · Attention · Retention · Feature fusion · Dynamic Areas of Interest

1 Introduction

With the rise of online and blended education, videos have become a key component of many courses, both in traditional schools and universities, as well as in newer learning settings such as MOOCs. Indeed, research in education has shown that videos can convey complex subjects in an effective and engaging manner, provided that best practices for video design are followed [20]. However, even well-designed videos can still be challenging for some students, and

teachers cannot easily detect these difficulties in remote settings. In particular, students' attention as well as information retention can fluctuate throughout a video due to both internal factors (e.g., tiredness, interest) and external factors (e.g., video length, difficulty) [10,21], which can hinder learning [8]. Hence, students might benefit from adaptive mechanisms that can monitor in real time how they watch educational videos and provide them with dedicated support as needed, e.g., [4,13]. Additionally, instructors and course designers could gain valuable insights into when and why attention and retention drop, allowing them to enhance both their videos and the overall course content accordingly.

Recent research has explored eye-tracking data to build machine learning (ML) models meant to detect students' attention while watching educational videos [1,6,12,25]. Eye tracking is well-suited for this task, because it can capture rich insights into how the students visually process video content. This task is typically performed by extracting a battery of statistical features over the different types of data generated by an eye-tracker (e.g., fixations, saccades, pupil size). Next, the features are fed into ML classifiers to predict binary labels of attention. This approach is relevant for the small datasets we usually deal with in educational research (e.g., a few hundred datapoints in [1,6,12,25]). Furthermore, standard eye-tracking features have been linked to cognitive and affective processes for decades [11], thus supporting interpretability. In our work, we leverage an educational video designed for a blended course on green chemistry to investigate how existing eye-tracking-based ML pipelines for predicting learners' attention could be enhanced by incorporating dynamic Areas of Interest (AOIs). To this end, we run a user study with 54 learners, and conduct an ML experiment on the data we collected. We also provide in open-source our eye tracking dataset and the code, to facilitate replication and enable cross-dataset analyses.

We aim to contribute to previous work in several ways. First, work on attention prediction in educational video viewing [1,6,12,25] has not extensively considered AOIs, an important component of eye tracking studies [11]. AOIs capture salient and meaningful parts of an application or media to analyze how users allocate their attention to the information on the screen and transition between them. This, in turn, could enable understanding students' learning strategies and engagement. In educational videos, only a few studies have considered broad AOIs (i.e., title, slides, speaker, most salient area), with no evidence that they can increase prediction performance [12,25]. Hence, building on this prior work, we investigate semantically richer AOIs capturing the type and importance of the fine-grained information displayed in the video, to understand whether such information can enhance both the prediction and the interpretability of ML models. We also report on the difficulty of extracting such AOIs in a dynamic video and the data processing pipeline we put in place to do so. Second, we extend previous work by exploring different approaches to fuse AOI-based features with other types of eye-tracking features (e.g., pupil size), to account for the heterogeneity of the underlying gaze processes. Third, beyond attention, we examine whether eye tracking can be used to predict information retention. To the best of our knowledge, no eye-tracking-based ML study has been conducted to predict

retention in educational videos. However, attention allocation does not always lead to information retention, and if such cases can be detected automatically, adaptive support could be designed to specifically address them. Altogether, we aim to provide a more comprehensive perspective on the value of dynamic and content-aware AOIs, combined with different feature fusion approaches, for improving attention and retention modeling in educational videos.

2 Related Work

There has been extensive work on predicting from eye gaze data attention loss and retention/comprehension during reading of educational and multimodal documents, e.g., [2,4,5,6,7]. In contrast, only a handful of studies have focused on educational videos to perform such predictions, specifically to predict mind wandering, a form of attention loss (see overview in [16]). In particular, Hutt et al. [12] explored the automatic detection of mind wandering using eye-tracking while watching a recorded lecture. To achieve this, they trained a Bayesian network to classify the attention levels of 32 learners at various moments in the video based on feature vectors derived from their gaze movements. They use a combination of global features (capturing students' gaze behaviors over the entire screen) and local features (capturing gaze on two AOIs: the most visually salient region of each frame, and the face of the lecturer). Their results show that their predictor outperforms random guessing, albeit with moderate performances (F1 of 0.47, vs. 0.3 for chance). They also found that AOIs hindered the model predictions, as opposed to using global features only. In a follow-up paper on the same lecture video, Bixler & D'Mello [6] were able to increase the prediction performance to an F1-score of 0.57, using a SVM classifier and global features only. Zhao et al. [25] explored the detection of mind wandering in two short MOOC's videos with 13 learners. They leveraged both global features, and local features defined for three AOIs (the speaker's face, the subtitles, and the lecture slides). They found that the best approach is to use a Naive Bayes classifier with global features, yielding a F1 score of 0.41, against a 0.29 baseline. Bühler et al. [1] focused on a different prediction task, namely predicting aware vs. unaware mind wandering while watching of a 60-minute lecture video with 87 learners. They used global features only and obtained their best performances with a SVM classifier (F1 of 0.33 vs. 0.24 for the baseline). These findings reveal that, while eye tracking can capture to some extent signals that are predictive of mind wandering better than chance, the performance of the trained classifiers remains overall moderate, highlighting the difficulty of this machine learning task. Other modalities for capturing attention during educational video watching include webcam recordings and EEG. Webcam recordings have shown limited predictive performance [22] and may raise ethical concerns due to the need to record students' faces. EEG has shown more promise in terms of prediction [3], but it requires wearing an intrusive headset and cannot indicate which part of the video the student is attending to, if any.

3 User Study

We describe the user study we conducted to investigate our research goals.

Stimulus. The video we used in our study is part of an introduction to green chemistry course. This course is delivered in a hybrid format, with the first part consisting of watching an introductory video about the "twelve principles" of green chemistry, as well as quizzes on these principles. The video was created by chemistry professors with the assistance of CAPSULE, the teaching and learning centre of Sorbonne University. The video lasts 7:30 minutes and was meant to follow best practices for online educational videos, in terms of length and content.

Participants. 54 students (38 females) participated in the study between January 2022 and March 2024. They were all undergraduate students in chemistry at Sorbonne University, and enrolled in the course on Green chemistry.

Protocol. The participants were first informed about the study and invited to sign a consent form. Next, they filled out a pre-survey (described below) and underwent a 9-point calibration with a Tobii Nano eye-tracker. We also measured their baseline pupil size by having them watch a white screen for 10 seconds. Next, the participants logged into the Moodle platform where the video is hosted. They were instructed to watch the video as they would at home, and that they were free to pause and seek within the video, as well as take notes. Participants watched the video alone in a room with constant lighting. This was to mimic as much as possible how such educational videos can be used by students. The average watch time was about 11 minutes (std. dev.=4.2 min). After having watched the video, students were asked to fill out a retention test and an attention survey (both described below). All collected data were kept fully anonymous. Following local regulations, the protocol and data collection plan were reviewed by a Data Protection Officer.

Material. In the *pre-survey*, participants were asked to provide some demographics, disclose their levels of tiredness, and self-report their knowledge of green chemistry on a 5-point Likert scale. They also answered three questions about green chemistry to gauge their existing knowledge on this topic.

The *retention test* is meant to capture what the participants retained from the video. It includes 15 multiple-choice questions, each targeting specific information (facts) provided in the video about green chemistry. We selected the 15 questions with a chemistry professor. Importantly, each item in the memory test corresponds to one specific part of the video, so that we can retrieve the eye-tracking data corresponding to that part. The 15 video parts corresponding to the 15 items were spaced about 20 to 25 seconds apart. In addition, for each question in the test, participants were also asked to report whether they already knew the answer to the question before watching the video. This was so that we could discard items that participants already knew beforehand and focus rather on what they actually retained from the video.

The *attention survey* is meant to capture the self-reported levels of attention of the students at specific parts of the video. Specifically, the survey includes 18 items, each composed of a screenshot from the video and a 5-point Likert scale to report the levels of attention when the screenshot was visible. Students could also

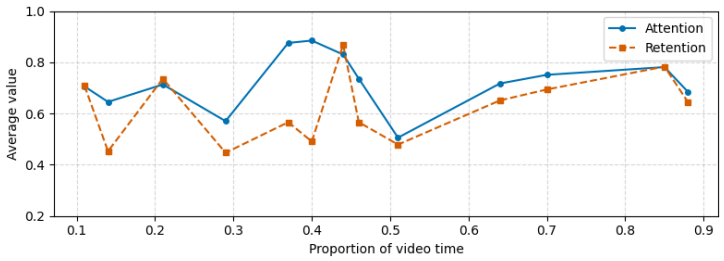

Fig. 1. Fluctuation in attention and retention throughout the video.

answer that they do not remember, and were in fact instructed to do so unless they were sure of their levels of attention. We used a retrospective survey because of the short length of the video, and because we can align and compare the outputs of the survey and the retention test. Another popular approach would have been probes, however, probes can be visually and aurally distracting and impact the learners' gaze behaviors, as well as their learning process, especially when the probe fires when they are taking notes.

As a result, each item in the attention and retention surveys can be mapped to one specific time within the video (called *item timestamp* from now on). Fig. 1 shows how the answers to both surveys (normalized from 0 to 1) evolve throughout the video. As a sanity check, this chart shows that the attention scores are overall higher than the retention ones, which makes sense given that attention is typically part of the memory process [19]. This provides further rationale for the need to study how to predict both attention and retention individually. Fig. 1 also indicates that attention peaked between 35% and 45% of the video, which is when the first principles of green chemistry (the main learning objective of the video) are defined.

4 Machine Learning Experiment

We use the eye-tracking data collected during the study to predict two binary targets, namely retention and attention. Retention is operationalized as whether participants answered the items in the retention test correctly. As said above, there were 15 items in this test, resulting in 15 items $\times$ 54 participants $=$ 810 binary labels (of which 39% indicate no retention). As for attention, it was operationalized based on the answers to the Likert-scale attention survey, which we discretized with a median split to distinguish between low and high attention. We target binary labels instead of the raw Likert-scale scores for attention as it is a common practice in related work (e.g., [2,4,14]) given that an adaptive system generally needs to decide should or should not receive adaptation, which makes binary classification best suited for this application. This results in 18 items $\times$ 54 participants $=$ 972 binary labels (of which 30% indicate low attention). In the experiment, we encoded the labels for low attention and lack of retention as "1",

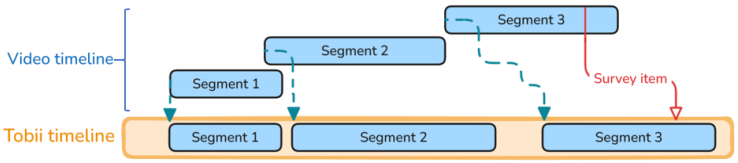

Fig. 2. Sample alignment between three video segments and the Tobii data steam.

and "0" otherwise, as our main goal is to predict when attention and retention drop for adaptation. We thus focus on these two binary classification tasks, and describe next how we process the eye-tracking data into feature vectors to build ML classifiers. The data, features and code are available in open-source at: `https://gitlab.lip6.fr/mocah-public/eye_tracking_green_chem_video`.

4.1 Data alignment

Because students were allowed to pause and seek within the video, as they would normally do at home, the timestamps in the Moodle video logs and in the Tobii eye-tracking data needed to be aligned. This was necessary to retrieve the actual eye-tracker timestamps matching the attention/retention item timestamps in the video. This was also required to define dynamic AOIs that only appear at specific times in the video, as discussed in the next subsection. The logs of the Moodle's video player (Panopto) are organized as successive *segments*, each indicating for how long a portion of the video was watched without interruption. These segments include a GMT timestamp, the start time in milliseconds (ms) of the segment within the video, and the watch duration in ms. A segment ends when the video is paused, reaches its end, or when the learners seek forward/backward. Unfortunately, we found out that the timestamps were inconsistent and often delayed by several seconds, which makes it very hard to align these logs with the eye-tracker ones. On the other hand, while the eye tracker records all learners' actions, it does not record the actual video times in case of seeking events. Thus, to fix the timestamps of the video logs and align them with the eye tracker's ones, we create a two-states (playing, paused) automaton that reads incrementally through both the video segments and eye-tracker action logs to find what specific learners' actions (playing, pausing, seeking) in the eye-tracker logs correspond to the start and end of a segment. Fig. 2 illustrates the result of this alignment process for three segments, where the learners sought backward between segments 1-2, then paused between segments 2-3.

This alignment process allows us to retrieve at what specific Tobii timestamp the information corresponding to an attention or retention item appeared. See for example the red solid line in Fig. 2 for a given survey item occurring during segment 3. Because segments can overlap in case of seeking events (e.g., segments 1 and 2 in Fig. 2), it is possible that an attention/retention item timestamp occurs in more than one segment. In those cases, it would be difficult to identify in which segments students were actually paying attention and/or retained infor-

Table 1. Number and summary of eye-tracking features.

a) Fixation (4) and Saccade features (15)
- Fixation rate, Fixation duration (mean, SD, Max)
- Saccade Duration, Distance, Velocity (mean, SD, max)
- Absolute and Relative saccade angles (mean, SD, rate)
b) Pupil width (10) and Eyes-screen distance features (6)
- Pupil width / eyes-screen distance (mean, SD, min, max)
- Pupil width / eyes-screen distance at first and last recorded fixation
- Pupil dilation velocity (mean, SD, min, max)
c) AOI Features (22 for AOI-2x1, 27 for AOI-imp, 72 for AOI-type)
- Fixation Rate, Fixation duration in AOI (mean, SD, max)
- Time to first fixation in AOI, Proportion of time in AOI
- Prop. of transitions from this AOI to every AOI
- Pupil width, pupil dilation velocity when looking at AOI (mean, SD, min, max)
- Eyes-screen distance when looking at AOI (mean, SD, min, max)

mation. Fortunately, this happened for only 8 retention labels and 12 attention labels, which we opted to remove.

4.2 Eye tracking data processing and feature sets

We use the EMDAT open-source Python library[1] to process the eye-tracking data. EMDAT produces a battery of eye-tracking features (listed in Table 1) specified over the entire display, and over specific AOIs. These features are commonly-used in eye tracking research [11].

Global features. These features, listed in Table 1a-b, are computed from gaze behaviors over the entire screen, namely: the user's fixations (gaze maintained at one point on the screen), saccades (quick eye movement between two fixations), pupil size (baseline-adjusted), and distance from both eyes to the screen (eye-screen distance, averaged over both eyes). These features are similar to research related to ours, e.g., [2,4,14,17]. Compared to them, we mainly add features related to pupil dilation velocity because of previous work linking them to reading comprehension and skill acquisition [18,23].

AOI features. In addition, we generate features for AOIs defined over specific parts of the video. These AOIs are meant to capture the specific gaze processing generated by each area, as well as gaze transitions between areas. To do so, we defined the pixel boundaries for three different sets of AOIs. The first one is a simple 2x1 grid with one area capturing the top part of the video where titles mostly lie, and the other one the rest (*AOI-2x1* from now on). This is meant to see whether a simple, static set of AOIs is sufficient. The second set captures six different areas based on the types of elements shown in the videos, namely titles, images, texts, mathematical formula, chemical synthesis/molecules, and data (*AOI-types* from now on). The third set captures three areas based on the

[1] `https://github.com/Human-AI-Interaction-UBC/EMDAT`

importance of these elements (high, medium, low) as annotated by the chemistry professor teaching the course (*AOI-imp* from now on).

For the second and third set, the AOIs had to be dynamic since the elements in the video only appear at specific time frames. In practice, EMDAT handles such dynamic AOIs by selecting only the eye-tracking data that fall within the AOI boundary at the timestamps during which the AOI is visible to compute the AOI features. To create the dynamic AOIs, we used the open-source VIA tool[2] to define the pixel boundaries of each element appearing in the video as well as their appearance and disappearance times. Each area was annotated with its type (for the AOI-types set) and levels of importance (for the AOI-imp set). Next, a Python script was created to parse the VIA's export, and the alignment process defined above was used to generate the series of timestamps in EMDAT format during which each AOI was actually visible. As a result, we obtained for each AOI the features defined in Table 1c. To the best of our knowledge, both this type of dynamic AOIs and the list of related features are novel in related work on attention detection in educational video.

Feature sets. We compute the above features for each retention and attention label, using 20 seconds intervals of eye-tracking data prior to the occurrence of each label. We chose 20 seconds because it loosely corresponds to the average gap durations in-between the survey items. It is also similar to the intervals used in related work, which typically range from 10 sec to 30 sec [1,6,12,25][3]. As a result, we obtained seven feature sets: (i) one with only global features and no AOI (*Overall-no-AOI feature set*); (ii); three feature sets with only the AOI-based features for each of the three sets of AOIs (*AOI-2x1-only set*, *AOI-type-only set*, and *AOI-imp-only set*); (iii) three feature sets composed of the global features augmented with each of the AOI feature sets (*Overall-AOI-2x1 set*, *Overall-AOI-type set*, and *Overall-AOI-imp set*). For feature fusion (described next), we also looked at four subsets of the *Overall-no-AOI set*, by separating the features related to fixations, saccades, pupils, and eye-screen distance, respectively.

4.3 Machine learning setup

We describe the process to build the ML classifiers for the two binary classification tasks (retention and attention) based on the aforementioned feature sets. In this work, we experiment with different approaches to fuse the different types of data obtained with the eye-tracker, i.e., fixations, saccades, pupil sizes, eye-screen distance, and AOIs. We do so because each of these data types present unique characteristics, and differ substantially in terms of number of features (see Table 1), frequency, units, and format. Borrowing from the literature on multimodal feature fusion, we leverage and compare three approaches, namely early, intermediate and late fusion [9,15], as described next.

Early fusion means concatenating all features in a single vector or tensor prior to training ML classifiers. This is the only approach used in the related

[2] VIdeo Annotator tool (VIA): `https://www.robots.ox.ac.uk/~vgg/software/via/`

[3] For completeness, we did try 10-sec and 30-sec intervals, but it did not have any substantial impact on the results.

work on attention detection in educational video [1,6,12,25]. While early fusion is the easiest to implement, it may lose some of the unique characteristics of smaller subsets of features. In our case, the seven feature sets described at the end of Section 4.2 are already the results of concatenating various eye-tracking features, and thus we simply leverage them to explore early fusion. Namely, for each of these seven feature sets and each of the two prediction tasks, we train a set of ML classifiers using the *scikit-learn* (`https://scikit-learn.org`) and *keras* (`https://keras.io`) Python libraries. We selected classifiers that have been shown to perform well on similar user modelling tasks and dataset sizes, without clear evidence as for which one is the best, e.g., [1,6,14,17,24,25], namely Random Forest (RF), SVM, Logistic Regression (LR) and Naive Bayes (NB). We also trained a sparse feed-forward neural network (FFNN1) composed of a fully connected layer with LeakyReLU activation, and dropout for sparsity so as to mitigate possible overfitting given our sample size. We also trained another network (FFNN2) with a second hidden layer, similar to the first one. Sigmoid activation is used in the final layer for binary prediction, alongside the Adam optimizer and the binary cross-entropy loss. To increase the interpretability of the network, we added a dot-product feature attention layer after the input layer, both to help the network focus on important features, and to be able to extract the attention weights for further investigation.

Late fusion means training a single classifier per type of features, and combining the trained classifiers via an ensemble mechanism. While it can miss fine-grained interaction between feature subsets, it allows to focus on each individual feature subset. To do so, we train the same classifiers as above using the exact same nested cross-validation approach, but this time for each of the four global features subsets (fixation, saccade, pupil, eye-screen distance), plus the three AOI-only sets (AOI-2x1-only, AOI-type-only, AOI-imp-only). We then train two ensemble classifiers with scikit-learn to combine the predictions from all of these classifiers, namely: a simple majority-voting classifier (EnsMV), and a decision tree. We use majority voting because this is the easiest and most straightforward ensemble technique. As for the decision tree (EnsDT), we use it to explore more complex relationships among the classifiers' outputs.

Intermediate fusion includes techniques to combine the different types of features as part of the inner working of a ML model. This approach is more complex to implement but provides a good compromise between early and late fusion. To do so, we leverage the high modularity of neural networks with keras. Specifically, we feed each of the four global feature subsets and the AOI-only sets into a separate fully connected layer, with LeakyReLU activation and dropout. We then concatenate the output of all of these layers into an intermediate one that we leverage for the same feature attention layer as described above for early fusion. The output is then provided to one or two fully connected layers similar to the FFNN1 and FFNN2 ones, respectively. As a result we obtain two networks that we label IFFNN1 and IFFNN2.

In total, we train for each prediction task a total of 42 classifiers for early fusion (7 feature sets x 6 classifiers), 14 classifiers for late fusion (7 feature sets x 2

Table 2. F1 for attention prediction. Bold means significantly better than the baseline.

Fusion	Model	Feature set						
		Overall-AOI-2x1	Overal-AOI-imp	Overall-AOI-type	AOI-2x1-only	AOI-imp-only	AOI-type-only	Overal-no-AOI
Early	FFNN1	0.31	0.55	**0.57**	0.33	**0.58**	**0.58**	0.31
	FFNN2	0.32	**0.58**	**0.57**	0.3	**0.58**	**0.59**	0.24
	LR	0.33	0.52	0.51	0.34	**0.58**	**0.58**	0.27
	NB	0.21	0.34	0.06	0.05	**0.56**	**0.57**	0.02
	RF	0.09	0.49	0.46	0.16	0.49	0.53	0.06
	SVM	0.01	0.43	0.16	0.07	0.52	0.54	0.04
Intermediate	IFFNN1	0.32	0.5	**0.58**	NA	NA	NA	0.30
	IFFNN2	0.33	0.52	**0.59**	NA	NA	NA	0.32
Late	EnsMV	0.2	0.47	0.39	0.18	0.51	0.5	0.12
	EnsDT	0.2	0.47	0.39	0.24	0.51	0.51	0.22
Baseline		0.44	0.44	0.4	0.44	0.41	0.44	0.43

ensemble classifiers), and 8 classifiers for intermediate fusion (2 classifiers for each AOI set and no-AOI). We build on top of these classifiers a random baseline per feature set and prediction task. All classifiers are trained and tested using nested, stratified 8-folds cross-validation (CV). At the inner loop of nested CV, invariant and highly correlated features are filtered out to ease the training process and avoid collinearity issues. Next, the features are normalized and univariate feature selection is used to reduce the dimension of the training data. Lastly, SMOTE is applied to balance the training sets and grid-search hyper-parameter tuning is conducted, still at the inner loop of CV on the training data only. The searched hyperparameter values are documented in the repository linked above. At the outer loop of CV, the performance of the trained models with the best found hyperparameter values is evaluated on a test set, and the process is repeated for each of the 8 folds. As for the performance metrics, we use the F1 score, similarly to previous work on attention detection [1,3,6,12,22].

5 Results

Table 2 provides the F1 scores averaged over the folds for the prediction of attention. It shows that the highest F1 score (0.59, against a 0.44 baseline) is reached by the IFFNN2 and FFNN2 classifiers with AOI-type. Table 3 shows the results for the prediction of retention, where the best F1 score (0.52, against a 0.44 baseline) is reached by the IFFNN1 and IFFNN2 classifiers with AOI-imp.

To explore these results, we first ascertain what classifiers are significantly better than the baselines. To do so, we run two linear mixed-effect models (one per prediction task) with the F1 score as the dependent variable, classifier and feature set as the factors, and the CV folds as the random effect. Both models

Table 3. F1 for retention prediction. Bold means significantly better than the baseline.

Fusion method	Model	Feature set						
		Overall-AOI-2x1	Overal-AOI-imp	Overall-AOI-type	AOI-2x1-only	AOI-imp-only	AOI-type-only	Overal-no-AOI
Early	FFNN1	0.28	0.47	0.38	0.4	0.47	0.45	0.35
	FFNN2	0.28	0.47	0.43	0.28	0.42	0.44	0.36
	LR	0.36	0.44	0.4	0.45	0.45	0.43	0.37
	NB	0.11	0.44	0.05	0.05	0.45	0.44	0.08
	RF	0.13	0.23	0.19	0.22	0.33	0.27	0.13
	SGB	0.19	0.28	0.24	0.23	0.34	0.34	0.26
	SVM	0.14	0.19	0.18	0.14	0.25	0.34	0.23
Intermediate	IFFNN1	0.37	0.5	0.45	NA	NA	NA	0.37
	IFFNN2	0.41	**0.52**	0.44	NA	NA	NA	0.38
Late	EnsMV	0.24	0.37	0.28	0.22	0.37	0.36	0.22
	EnsDT	0.28	0.41	0.35	0.25	0.43	0.41	0.26
Baseline		0.41	0.42	0.42	0.44	0.42	0.44	0.44

reveal a significant interaction effect ($p<0.001$) between classifier and feature set, which is consistent with the substantial differences in F1 across the feature sets in Tables 2-3. To investigate these interaction effects, we run post-hoc t-test pairwise comparisons with the Benjamini and Hochberg adjustment for all combination of classifier and feature set. We report in bold in Tables 2-3 the models for which the pairwise comparisons show a significant improvement ($p<0.05$) over the baseline after adjustment.

For attention (Table 2), several models significantly outperformed the baseline, with no significant difference among them. Interestingly, all of them leveraged either the AOI-imp or AOI-type sets. No model trained without AOI (Overal-no-AOI) or with the static AOI-2x1 set could outperform the baseline. This indicates the value of leveraging these AOI sets for our prediction task. In terms of performance, while the feed-forward neural network generally performed well with both early and intermediate fusion, reaching F1 of 0.58 to 0.59, we also found that more traditional AI models (LR and NB) can reach very similar performances with F1 of up to 0.58. This suggests that there may be simpler, linear relationships among some of the gaze features and attention, which can be captured even by simple classifiers. No model with late fusion outperformed the baseline, indicating that this approach for feature fusion is not effective. While it is difficult to draw formal comparison with prior work given the difference in dataset, our classifiers' F1 scores are substantially higher than most of the prior results reported in Section 2, and even slightly better than the most successful approach so far (F1 of 0.57 in [6]).

For retention, only one model turned out to be significantly better than the baseline after adjustment: IFFNN2 with AOI-imp. While the performance of this model (F1 of 0.52) remains moderate, it does suggest that intermediate fusion

and AOI-imp can capture some information from the eye-tracking data that can to some extent capture retention. These results also show that predicting retention from eye-tracking is much more challenging than predicting attention.

6 Discussion and conclusion

In this work, we investigated the value of eye tracking for predicting learners' attention and information retention (memory) while watching educational videos. To this end, we collected an eye-tracking dataset in a blended course on green chemistry where enrolled students were required to watch an introductory video. We then leveraged the collected data to train machine learning classifiers to predict binary labels of attention and of retention.

Our main focus was to explore the value of dynamic AOIs to improve the performance of the classifiers and provide insights into specific gaze behaviors within the video. Our results do show that dynamic AOIs capturing either the importance or type of information provided in the video, can substantially enhance the performance of ML classifiers. We provide a pipeline for leveraging such AOIs, so as to facilitate future research. While we expect AOI features to provide insights into the underlying gaze processes related to the video content, there is also a need to investigate this aspect with instructors. We also focus on examining several feature fusion approaches, to take advantage of the fact that eye trackers typically output various types of data. We found that for retention prediction, only a neural network with intermediate fusion was able to outperform the baseline, highlighting the potential of this approach. We hypothesize that this approach could be even more powerful by fusing other modalities (e.g., interaction data, physiological sensors, as in [1]), which could be investigated in the future. In general, we found that predicting retention is more challenging than predicting attention, suggesting another key direction for future research.

In future work, we plan to study automatic approaches to define the type and importance of AOIs, to further facilitate AOI integration. We also aim to replicate our work with other videos, to study the generalizability of our research. Ultimately, our long-term goal is to leverage our classifiers in real time to drive real-time support when learners exhibit low predicted levels of attention/retention, and inform instructors about the learners' difficulties, in order to increase learners' active engagement, such as taking notes, pausing to reflect, and applying their knowledge, rather than passive watching.

Acknowledgments. This work was partially supported by the Sorbonne Center for Artificial Intelligence (SCAI) via a student scholarship. We thank CAPSULE and the *"Label vert"* at Sorbonne for their support with the user study.

References

1. Bühler, B., Bozkir, E., Deininger, H., Goldberg, P., Gerjets, P., Trautwein, U., Kasneci, E.: Detecting aware and unaware mind wandering during lecture viewing: A

multimodal machine learning approach using eye tracking, facial videos and physiological data. In: Proceedings of the 26th International Conference on Multimodal Interaction. pp. 244–253 (2024)
2. Copeland, L., Gedeon, T., Caldwell, S.: Effects of text difficulty and readers on predicting reading comprehension from eye movements. In: 2015 6th ieee international conference on cognitive infocommunications (coginfocom). pp. 407–412. IEEE (2015)
3. Dhindsa, K., Acai, A., Wagner, N., Bosynak, D., Kelly, S., Bhandari, M., Petrisor, B., Sonnadara, R.R.: Individualized pattern recognition for detecting mind wandering from eeg during live lectures. PloS one **14**(9), e0222276 (2019)
4. D'Mello, S., Kopp, K., Bixler, R.E., Bosch, N.: Attending to attention: Detecting and combating mind wandering during computerized reading. In: Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems. pp. 1661–1669 (2016)
5. D'Mello, S.K., Southwell, R., Gregg, J.: Machine-learned computational models can enhance the study of text and discourse: A case study using eye tracking to model reading comprehension. Discourse processes **57**(5-6), 420–440 (2020)
6. E. Bixler, R., K. D'Mello, S.: Crossed eyes: Domain adaptation for gaze-based mind wandering models. In: Acm symposium on eye tracking research and applications. pp. 1–12 (2021)
7. Faber, M., Bixler, R., D'Mello, S.K.: An automated behavioral measure of mind wandering during computerized reading. Behavior Research Methods **50**, 134–150 (2018)
8. Faber, M., Krasich, K., Bixler, R.E., Brockmole, J.R., D'Mello, S.K.: The eye–mind wandering link: Identifying gaze indices of mind wandering across tasks. Journal of experimental psychology: human perception and performance **46**(10), 1201 (2020)
9. Guarrasi, V., Aksu, F., Caruso, C.M., Di Feola, F., Rofena, A., Ruffini, F., Soda, P.: A systematic review of intermediate fusion in multimodal deep learning for biomedical applications. arXiv preprint arXiv:2408.02686 (2024)
10. Hollis, R.B., Was, C.A.: Mind wandering, control failures, and social media distractions in online learning. Learning and Instruction **42**, 104–112 (2016)
11. Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., Van de Weijer, J.: Eye tracking: A comprehensive guide to methods and measures. oup Oxford (2011)
12. Hutt, S., Hardey, J., Bixler, R., Stewart, A., Risko, E., D'Mello, S.K.: Gaze-based detection of mind wandering during lecture viewing. International Educational Data Mining Society (2017)
13. Hutt, S., Krasich, K., R. Brockmole, J., K. D'Mello, S.: Breaking out of the lab: Mitigating mind wandering with gaze-based attention-aware technology in classrooms. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–14 (2021)
14. Jaques, N., Conati, C., Harley, J.M., Azevedo, R.: Predicting affect from gaze data during interaction with an intelligent tutoring system. In: Intelligent Tutoring Systems: 12th International Conference, ITS 2014, Honolulu, HI, USA, June 5-9, 2014. Proceedings 12. pp. 29–38. Springer (2014)
15. Jiao, T., Guo, C., Feng, X., Chen, Y., Song, J.: A comprehensive survey on deep learning multi-modal fusion: Methods, technologies and applications. Computers, Materials & Continua **80**(1) (2024)
16. Kuvar, V., Kam, J.W., Hutt, S., Mills, C.: Detecting when the mind wanders off task in real-time: An overview and systematic review. In: Proceedings of the 25th international conference on multimodal interaction. pp. 163–173 (2023)

17. Lallé, S., Murali, R., Conati, C., Azevedo, R.: Predicting co-occurring emotions from eye-tracking and interaction data in metatutor. In: International Conference on Artificial Intelligence in Education. pp. 241–254. Springer (2021)
18. Martínez-Gómez, P., Aizawa, A.: Recognition of understanding level and language skill using measurements of reading behavior. In: Proceedings of the 19th international conference on Intelligent User Interfaces. pp. 95–104 (2014)
19. Mayer, R.: Cognitive theory of multimedia learning (2005)
20. Noetel, M., Griffith, S., Delaney, O., Sanders, T., Parker, P., del Pozo Cruz, B., Lonsdale, C.: Video improves learning in higher education: A systematic review. Review of educational research **91**(2), 204–236 (2021)
21. Risko, E.F., Anderson, N., Sarwal, A., Engelhardt, M., Kingstone, A.: Everyday attention: Variation in mind wandering and memory in a lecture. Applied Cognitive Psychology **26**(2), 234–242 (2012)
22. Stewart, A., Bosch, N., Chen, H., Donnelly, P., D'Mello, S.: Face forward: Detecting mind wandering from video during narrative film comprehension. In: International conference on artificial intelligence in education. pp. 359–370. Springer (2017)
23. Toker, D., Lallé, S., Conati, C.: Pupillometry and head distance to the screen to predict skill acquisition during information visualization tasks. In: Proceedings of the 22nd International Conference on Intelligent User Interfaces. pp. 221–231 (2017)
24. Veliyath, N., De, P., Allen, A.A., Hodges, C.B., Mitra, A.: Modeling students' attention in the classroom using eyetrackers. In: Proceedings of the 2019 ACM Southeast Conference. pp. 2–9 (2019)
25. Yue, Z., Lofi, C., Hauff, C.: Scalable mind-wandering detection for moocs: A webcam-based approach [c]. In: European Conference on Technology Enhanced Learning. Springer, Cham. Tallinn, Estonia. pp. 330–344 (2017)