Comparison of Eye-Tracking Representations for Engagement Inference in Knowledge Workers

Tomi Božak
Jožef Stefan Institute
Ljubljana, Slovenia
tb85088@student.uni-lj.si

Zoja Anžur
Jožef Stefan Institute
Ljubljana, Slovenia
zoja.anzur@ijs.si

Gašper Slapničar
Jožef Stefan Institute
Ljubljana, Slovenia
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
gasper.slapnicar@ijs.si

Abstract

Eye tracking is an appealing sensor for inferring momentary psychological states at work. Before training a model, a practitioner must decide how to represent a gaze window. We compared four representation families on identical three-second windows drawn from knowledge workers across 657 full-workday recordings: raw multichannel signals (pupil size, gaze position and eye–screen distance), handcrafted ocular features, and embeddings obtained from pretrained eye-tracking (GazeMAE) and time-series (MOMENT) encoders, both used frozen. The target was a single self-reported item measuring momentary engagement. All four were evaluated under leave-one-subject-out cross-validation with per-fold preprocessing, using a random forest and an MLP classifier against a majority baseline. Raw signals gave the highest observed performance on the left-out participant (0.536 balanced accuracy), handcrafted features followed closely overall, and neither frozen encoder improved on the raw signal. Absolute discrimination was weak throughout (best ROC-AUC 0.553) and between-participant variability exceeded the differences between representations, so we report the ranking as preliminary, cost-aware guidance.

Keywords

eye tracking, engagement inference, representation learning, GazeMAE, MOMENT, subject generalisation

1 Introduction

Interactive systems that adapt to the person using them need evidence about that person's momentary psychological state. In knowledge work, momentary engagement, particularly immersion in the task at hand, is an attractive target. It fluctuates within a working day and is only weakly visible in application logs. Eye tracking is a plausible sensor for estimating it [7, 9]. Gaze position, fixation and saccade dynamics, blinks and pupil size are available continuously and non-intrusively from a screen-mounted tracker and require no action from the user.

Turning a stream of gaze samples into model input forces a design decision that is rarely examined on its own, even though benchmark evaluations in time-series classification show that representation families differ substantially in accuracy on the same data [8]. A window of eye-tracking data can be presented to a classifier in at least four ways. The raw multichannel signal uses selected tracker channels directly, at the price of high dimensionality and little domain structure. Handcrafted ocular features encode domain knowledge and remain interpretable, while learned representations replace manual feature design with embeddings from pretrained encoders [2, 3, 4, 5]. Here we compare two such encoders: gaze-specific GazeMAE and general-purpose time-series MOMENT [1, 6]. The four options differ substantially in preprocessing effort, storage and compute, yet they are seldom compared under the same conditions: published results differ in windows, labels, classifiers and, most consequentially, in validation protocol, so an apparent advantage of one representation can just as easily come from an optimistic split [10]. We therefore ask which eye-tracking representation supports the most reliable inference of self-reported momentary engagement in unseen knowledge workers. We conducted a controlled comparison of raw signals, handcrafted features, GazeMAE embeddings and MOMENT embeddings on identical three-second windows, an identical binary engagement target and identical leave-one-subject-out folds. Both pretrained encoders were kept frozen and only a downstream classifier was trained, which is the out-of-the-box setting a practitioner would try first.

2 Related Work

Eye tracking has long been used as behavioural evidence for internal user states. Pupil dilation is a classical correlate of processing load [7], and fixation, saccade, blink and pupil statistics computed over short windows remain common inputs to models of workload, affect and engagement [2, 5]. Our own prior work illustrates the appeal and the limits of this pipeline – handcrafted ocular features support coarse affective distinctions [4]. A systematic evaluation across several prediction targets also shows that task type is recognised far more reliably than graded workload, and that reported accuracy depends strongly on how the prediction target and the validation split are defined [3].

Learned representations offer an alternative. GazeMAE pretrains temporal autoencoders on raw position and velocity signals and shows that the resulting gaze-specific embeddings support downstream classification through a simple linear head [1]. MOMENT takes a broader approach by pretraining a multichannel encoder for transfer across time-series domains without task-specific pretraining [6]. The value of both encoders for modest, subject-dependent recordings of spontaneous behaviour remains unclear.

A further concern is evaluation design. Windows drawn from the same person are highly dependent, so random window splits and preprocessing fitted on the full dataset inflate reported performance, an effect documented in detail for other biosignal domains [10]. Comparisons across representations are interpretable only when every candidate is evaluated on identical windows and

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.skui.85

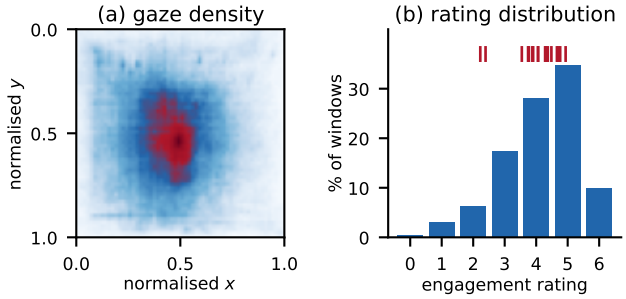

Figure 1: Descriptive context for the modelling cohort. (a) Gaze-coordinate density on the normalised display, pooled over all on-screen samples. (b) Distribution of the engagement rating over the 140,531 modelled windows. The red ticks mark the 17 individual participant means, which span 2.2–4.9 and motivate the participant-aware construction of the target.

labels under a subject-disjoint protocol with per-fold preprocessing.

3 Methodology

3.1 Data and Eye-Tracking Representations

The data came from a study conducting multimodal data collection, in which knowledge workers were recorded during their daily routine at their own workplace. The data being collected includes RGB video, eye tracking, microphone audio and application usage. This paper uses the eye-tracking modality only. Data collection is ongoing and the dataset is not yet public. This preliminary study used the cohort available at the time of writing, which includes 17 participants and 657 full-workday recordings, with 19 to 79 recordings per participant (median 32). The tracker sampled at 60 Hz, and we retained five channels: left and right pupil size, gaze x and y, and the average eye–screen distance.

Labels came from an experience-sampling protocol, in which participants rated their momentary state in situ rather than recalling it afterwards. Participants were prompted several times per working day with a short questionnaire covering engagement, affect, fatigue, stress, physical comfort and workplace interactions. Our target was the immersion item, "Right now I am immersed in my work", one of three momentary engagement items answered at every prompt and rated on a seven-point scale from 0 to 6. We refer to it as the *engagement rating*. Each prompt labelled the 15 minutes of recording that preceded it, a labelling horizon we chose ourselves, and only those intervals entered this study.

Recordings were segmented into fixed three-second windows with a 2.25-second step (25 % overlap). The four representation exports were joined on a per-window identifier, so all models saw exactly the same windows. At 60 Hz, each window was expected to contain 180 frames, although sampling gaps sometimes reduced this number. A window was kept only if at least 30 % of its recorded frames were valid and at least 32 valid frames were available. These permissive thresholds accommodated frequent tracking loss in unconstrained office work. The 32-frame minimum corresponds to about 0.53 seconds at 60 Hz, which can include two fixations of typical duration [9]. Gaps inside retained windows were linearly interpolated. Of roughly 7.3 million recorded windows, only those inside a labelled interval are usable. Handcrafted features were available for 177,244 of them, 157,685 of those also had both encoder embeddings, and 140,531 carried a numeric engagement rating. The final modelling cohort accounts for about 2 % of all recorded windows and contains between 2,136 and 25,447 windows per participant.

Display resolutions ranged from 1920×1080 to 3840×2160 pixels and differed between participants, and one participant used two resolutions across recordings. For the raw representation, gaze coordinates were normalised to $[0, 1]$ using each source recording's declared resolution, and out-of-screen coordinates were retained. Overall, 94.7 % of samples fell on the display, and only those entered the descriptive density map in Figure 1(a), which shows the expected central concentration together with broad coverage of the screen.

Four representations were computed from these windows. Coordinates are display-normalised for the raw export and in visual-angle units for GazeMAE; the handcrafted features and MOMENT embeddings retain pixel coordinates. (1) **Raw signals** use the five channels, each linearly resampled to 120 samples so that the whole window is covered, flattened into a 600-dimensional vector. (2) **Handcrafted features** describe a window through the ocular events and signals it contains, following the feature set of our earlier eye-tracking work [3, 4]. Offline detection used missing-gaze intervals for blinks, I-DT for fixations and non-blink intervals between fixations for saccades. These events contribute count, rate and duration statistics, alongside fixation location and dispersion and saccade amplitude and velocity. The remaining groups summarise pupil size and its rate of change, the eye–screen distance and its dynamics, and missing-data counts and fractions for selected signals. This gives 181 numeric features per window, which were filtered inside each fold as described in Section 3.2. Window identifiers, timestamps and questionnaire columns were not used as model input. (3) **GazeMAE embeddings** were computed with the public micro–macro autoencoder checkpoints applied to gaze position and velocity resampled to 500 Hz, with the 128-dimensional position and velocity codes concatenated into 256 dimensions [1]. Following the encoder's own preprocessing, coordinates were rescaled to 35 pixels per degree of visual angle using each recording's display resolution and the measured eye–screen distance. (4) **MOMENT embeddings** were calculated with the MOMENT-1-LARGE encoder applied to the five-channel windows, with an input length of 512 and mean pooling, giving 1,024 dimensions [6]. Its encoder standardises each channel per window, so the input coordinate scale is immaterial. Both pretrained encoders were used out of the box and kept frozen, with no weight updates.

3.2 Experimental Design

Engagement ratings were strongly person-specific, with participant means spanning 2.2 to 4.9 on the 0–6 scale (Figure 1(b)). An absolute threshold would therefore mostly encode participant identity. Inside every fold, the ratings of each training participant were mean-centred separately. The held-out participant's ratings were centred with the unweighted mean of the training participants' means, which avoids using held-out labels to set the threshold. A window is positive when its centred rating exceeds zero. Across the left-out folds, 81,341 of the 140,531 windows (57.9 %) were positive, and the participant-specific rates ranged from 0 % to 95 %.

The primary protocol was subject-disjoint cross-validation, with one of the 17 participants held out in each fold. Everything data-dependent was fitted inside the fold. Handcrafted-feature

Table 1: Leave-one-subject-out results for binary engagement inference, averaged over held-out participants. Balanced accuracy is the primary metric; its chance level is 0.500 by construction, Δ is its paired improvement over the majority baseline and SD is its spread across participants. M-F1 is macro-F1, AUC is ROC-AUC and Acc. is plain accuracy. RF is the random forest. Dimensionality is 600 for raw, 58–60 for handcrafted, 256 for GazeMAE and 1,024 for MOMENT. All metrics but accuracy use the 16 folds whose test participant is not single-class. Best value per metric column in bold.

	Bal. acc.	Δ	SD	M-F1	AUC	Acc.
Majority baseline	.500	—	.000	.326	.500	.524
Raw, RF	**.536**	**+.036**	.075	.464	.550	.528
Raw, MLP	.534	+.034	.050	**.472**	**.553**	**.542**
Handcrafted, RF	.508	+.008	.035	.418	.527	.528
Handcrafted, MLP	.521	+.021	.032	.451	.535	.531
GazeMAE, RF	.510	+.010	.027	.441	.514	.540
GazeMAE, MLP	.505	+.005	.031	.447	.508	.532
MOMENT, RF	.501	+.001	.013	.380	.504	.521
MOMENT, MLP	.497	−.003	.010	.362	.504	.507

filtering, median imputation and standardisation were all fitted on the training participants only and then applied with unchanged parameters to the held-out participant. The filtering drops columns with more than 60 % missing values, zero variance, or absolute Pearson correlation above 0.8 with a retained column, leaving 58–60 of the 181 feature candidates.

Three predictors were run on every representation: a majority baseline, a random forest (300 trees, balanced class weights) and an MLP (1028–512–256–128–64 units, GELU activations, layer normalisation, Adam with learning rate 2×10^{-4} and weight decay 10^{-3}, batch size 1024, at most 200 epochs with early stopping after 15 epochs without improvement). For the MLP, a stratified 20 % split of the outer training windows was used for early stopping. A single seed was used throughout.

We report balanced accuracy and its paired improvement over the majority baseline as primary measures, because the positive rate varies strongly between participants, plain accuracy partly measures a participant's own base rate, while balanced accuracy has a fixed chance level of 0.500 that makes folds comparable. Plain accuracy, macro-F1 and ROC-AUC are reported as secondary measures. All metrics were aggregated as unweighted means over participants. One participant self-reported every window below the training centre, so that test fold contains a single class. Balanced accuracy, macro-F1 and ROC-AUC are therefore defined for 16 folds and accuracy for all 17. As a separate within-participant check, we also ran a one-step persistence predictor within recordings, carrying the last observed label forward in 90-window test blocks after at least 300 calibration windows.

4 Results

Table 1 reports the main comparison, in which every pairing of representation and model sat close to chance. Raw signals gave the highest observed result with both classifiers: 0.536 balanced accuracy with the random forest, 0.036 above the 0.500 chance level, and 0.534 with the MLP, which led on macro-F1 (0.472), ROC-AUC (0.553) and plain accuracy (0.542). Handcrafted features followed, reaching 0.521 balanced accuracy with the MLP and improving on chance with both classifiers. GazeMAE stayed just above chance, at 0.510 balanced accuracy with the random forest and 0.505 with the MLP. MOMENT stayed at chance with both classifiers.

Handcrafted features ranked second on every metric except plain accuracy, where GazeMAE was second. Across models, balanced accuracy differed from chance by −0.003 to +0.036. For the leading raw result, the 0.075 standard deviation across participants exceeded both its 0.036 margin over chance and the 0.015 gap to the best handcrafted result. On balanced accuracy both classifiers placed raw first and MOMENT last, while the two middle representations swapped; within each representation their difference was smaller than that spread, so we do not read a ranking between them.

Figure 2 shows the same difficulty geometrically. None of the four representations arranges windows by engagement rating, and the colours are as mixed in the learned embeddings as in the raw signal. Colouring the identical projections by participant is more informative. Raw signals form clear participant-dominated patches, handcrafted features and GazeMAE group more weakly, and the MOMENT projection is close to uniform. The clearest visible structure is therefore person-specific, and the weaker grouping in the frozen embeddings may reflect smoothing by the encoders. Because the projections are exploratory, this interpretation remains a conjecture.

As a check on the label series, a one-step persistence predictor was evaluated over 1,434 temporal blocks. Participant-averaged accuracy reached 0.999, while balanced accuracy reached 0.960 and macro-F1 reached 0.965 over the 13 participants with two-class test blocks. Each questionnaire answer is broadcast unchanged to every window of the preceding 15-minute interval (Section 3), so most consecutive window pairs share a label by construction. Because the predictor uses observed labels within participants, its result cannot be compared directly with the left-out-participant results above.

5 Discussion

On this dataset, raw signals gave the highest observed scores among the four representations for binary engagement inference in unseen knowledge workers, with handcrafted features providing the strongest overall alternative and no improvement from the two frozen embeddings. The small margins over chance show that momentary self-reported engagement is difficult to predict across people from three-second gaze windows collected during unconstrained, real-world office work, regardless of representation.

We interpret the lack of improvement from the frozen encoders as a mismatch between their pretraining data and this task. GazeMAE was pretrained on controlled stimulus-viewing corpora and MOMENT on generic time-series benchmarks [1, 6]. Neither was pretrained for day-long, spontaneous office gaze with its dropouts, idle stretches and heterogeneous displays. These results concern frozen transfer only, and the next step is to fine-tune the encoders on office gaze, starting with GazeMAE, which came closest to the handcrafted baseline (Section 4).

For practical use, raw signals offer the simplest input pipeline, requiring no feature engineering, pretrained checkpoint or encoder inference pass. Handcrafted features should still be considered because they remain interpretable, and our earlier feature-based analyses found them effective on related eye-tracking targets [3, 4]. Our results do not support a frozen general-purpose encoder as the default choice.

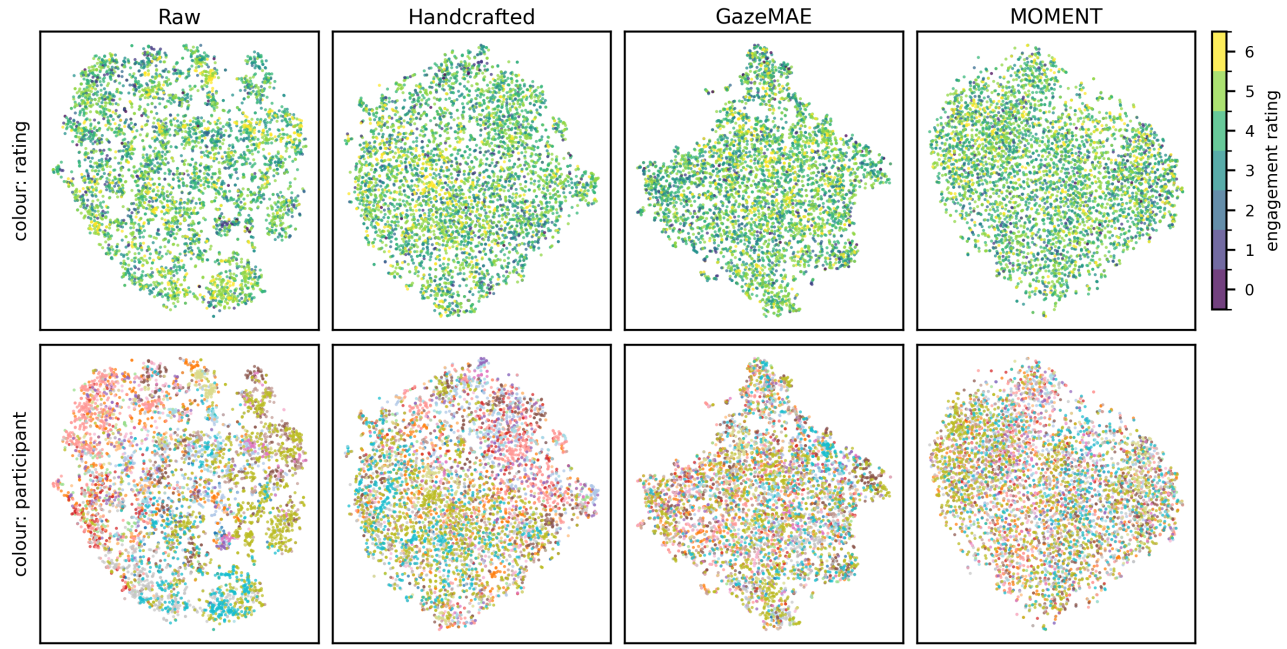

Figure 2: t-SNE projections of the four representations over a common sample of 5,000 windows, coloured by engagement rating (top) and by participant (bottom). Participant colours form visible patches in the raw projection and mix almost uniformly in the MOMENT projection, while no projection shows regions that follow the engagement rating.

Several limitations bound these conclusions: the study uses one unpublished and still-growing dataset with 17 participants. The target is one noisy self-report assigned retrospectively to a 15-minute interval and binarised with a per-fold centring rule. Overlapping windows and repeated labels within each interval reduce the effective sample size. The results depend on one windowing scheme, one seed, one set of quality-control rules and two specific public checkpoints. Future work should validate the comparison on further datasets, explore calibrated personalisation for new users, and test longer windows and aggregated targets that match the granularity at which engagement is actually reported.

6 Conclusion

We compared raw multichannel windows, handcrafted ocular features and frozen GazeMAE and MOMENT embeddings under one identical, leakage-safe leave-one-subject-out protocol for momentary engagement inference in knowledge workers. Raw windows produced the highest observed result with 0.536 balanced accuracy and 0.553 ROC-AUC on left-out participants, and out of the box the frozen encoders did not improve on them, despite their higher preprocessing and compute cost. For practitioners with limited labelled data, raw windows or handcrafted features are the better starting point, while frozen encoders need domain-specific validation or adaptation before use.

Acknowledgements

This work was supported by the bilateral Weave project, funded by the Slovenian Agency of Research and Innovation (ARIS) under grant agreement N1-0319 and by the Swiss National Science Foundation (SNSF) under grant agreement 214991.

The authors acknowledge the use of OpenAI Codex with GPT-5.6 Sol and Anthropic Claude Code with Claude Opus 5 for generating text suggestions during the preparation of this paper. The authors reviewed and edited all suggestions to ensure their accuracy and relevance to the research. The authors take full responsibility for the contents of the paper.

References

[1] Louise Gillian C. Bautista and Prospero C. Naval Jr. 2021. GazeMAE: general representations of eye movements using a micro-macro autoencoder. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 7004–7011. doi:10.1109/ICPR48806.2021.9412761.

[2] Robert Bixler and Sidney D'Mello. 2016. Automatic gaze-based user-independent detection of mind wandering during computerized reading. *User Modeling and User-Adapted Interaction*, 26, 1, 33–68. doi:10.1007/s11257-015-9167-1.

[3] Tomi Božak, Shivalika Goyal, Marc Langheinrich, Martin Gjoreski, and Gašper Slapničar. 2026. Evaluating feature-based machine-learning models with post hoc explainability for eye-tracking-based task type and workload inference. *AI*, 7, 8, 325. doi:10.3390/ai7080325.

[4] Tomi Božak, Mitja Luštrek, and Gašper Slapničar. 2024. Feature-based emotion classification using eye-tracking data. In *Proceedings of the 27th International Multiconference Information Society (IS 2024), Volume A: Slovenian Conference on Artificial Intelligence*. Ljubljana, Slovenia. https://is.ijs.si/?p=16837.

[5] Lex Fridman, Bryan Reimer, Bruce Mehler, and William T. Freeman. 2018. Cognitive load estimation in the wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–9. doi:10.1145/3173574.3174226.

[6] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. 2024. MOMENT: a family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 16115–16152. https://proceedings.mlr.press/v235/goswami24a.html.

[7] Daniel Kahneman and Jackson Beatty. 1966. Pupil diameter and load on memory. *Science*, 154, 3756, 1583–1585. doi:10.1126/science.154.3756.1583.

[8] Matthew Middlehurst, Patrick Schäfer, and Anthony Bagnall. 2024. Bake off redux: a review and experimental evaluation of recent time series classification algorithms. *Data Mining and Knowledge Discovery*, 38, 4, 1958–2031. doi:10.1007/s10618-024-01022-1.

[9] Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124, 3, 372–422. doi:10.1037/0033-2909.124.3.372.

[10] Matthew Rosenblatt, Link Tejavibulya, Rongtao Jiang, Stephanie Noble, and Dustin Scheinost. 2024. Data leakage inflates prediction performance in connectome-based machine learning models. *Nature Communications*, 15, 1, 1829. doi:10.1038/s41467-024-46150-w.