EEG-to-Scanpath: A Unified Model to Decode Eye Movements from EEG

Ard Kastrati[1]* **Radim Urban**[1]* **Martyna Beata Plomecka**[2]
Ruben Widmer[1] **Roger Wattenhofer**[1] **Nicolas Langer**[2]

[1]ETH Zurich, Switzerland
[2]University of Zurich, Switzerland

Abstract

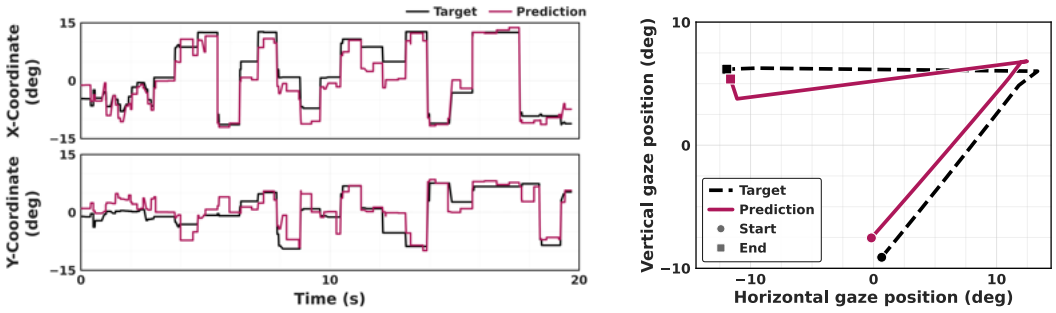
Eye movement-related activity in EEG is typically discarded as an artifact, obscuring a potentially rich source of behavioral information. Here, we show that this activity can instead be used to directly reconstruct scanpaths from EEG. We present an end-to-end unified model that predicts the timing and spatial properties of fixations, saccades, and blinks in a single forward pass. We train the model on existing publicly available datasets and introduce a new dataset recorded across diverse tasks with two EEG hardware setups. Our results show that the model generalizes across participants, paradigms, and hardware, and can be rapidly adapted to new users using only five minutes of calibration data. This work demonstrates that EEG contains recoverable gaze information, enabling scanpath reconstruction in legacy EEG datasets and clinical settings where conventional eye tracking is impractical.

1 Introduction

Eye movements offer a moment-to-moment window into attention and perception, making eye tracking a crucial method in psychology, neuroscience, and human–computer interaction [1, 2]. In many experimental settings, eye tracking is combined with electroencephalography (EEG), forming what is often considered an "ideal" setup for studying brain–behavior relationships [3, 4, 5]. Eye movements provide a direct measure of attention, enabling researchers to verify task compliance, quantify visual exploration strategies, and derive higher-level behavioral features such as fixation/saccade statistics or scanpath structure [6]. Crucially, many of these applications do not require sub-degree spatial accuracy. Instead, coarse but structured representations of visual behavior are often sufficient. For example, in brain–computer interface (BCI) settings, predictive models frequently rely on high-level gaze features rather than precise point-of-gaze estimation [7]. In addition, gaze signals are commonly used for quality control, with trials excluded when participants fail to maintain fixation, look away from the screen, or exhibit blinks and signs of drowsiness or cognitive load [4, 8].

Today, infrared video-based eye trackers are the most common approach in these scientific studies. Although hardware costs are gradually becoming less prohibitive, running concurrent EEG and eye-tracking experiments remains technically challenging. It requires operator training, time-consuming physical setup, calibration for each session [9], and precise hardware synchronization [2]. In many scenarios, including mobile or clinical settings and studies involving infants or low-vision populations [10], reliable eye tracking is difficult or infeasible. Furthermore, even when deployed, eye trackers can fail due to occlusions, such as eyewear or medical equipment, poor lighting conditions, or partial eye closure, as during drowsiness or sleep. As a result, a large fraction of neurophysiological datasets are collected using EEG alone, without accompanying gaze annotations, and even multimodal recordings frequently suffer from dropouts [11].

*These authors contributed equally to this work. Corresponding author: Ard Kastrati, akastrati@ethz.ch



(a) Decoded scanpath in X and Y coordinates over time

(b) Decoded scanpath in screen coordinates

Figure 1: **EEG-to-scanpath decoding.** Given EEG signals as input, our model reconstructs the corresponding sequence of eye movements (scanpaths), including fixations, saccades and blinks. The left panel shows decoding of an example from the test set in the time domain. The right panel shows the reconstructed scanpath in the screen from another example from test set. The model captures both temporal dynamics and spatial gaze behavior.

At the same time, neuroscience research has long established that eye movements generate strong electrical potentials that are naturally recorded by EEG electrodes [12]. Signals originating from the frontal eye fields, the visual cortex, the eye's natural electrical charge, and the surrounding muscles contain rich information about where and when the eyes move. Yet standard neurophysiological pipelines have typically treated this activity simply as an artifact to be removed. Techniques such as Independent Component Analysis (ICA) are routinely applied to filter out and discard these eye movements so that the underlying brain waves can be studied in isolation.

A primary reason eye movements are treated as an artifact rather than as a useful signal for extracting behavioral information is the significant noise inherent in EEG recordings. In this work, we demonstrate that these signal-to-noise challenges can be overcome and that this activity can instead be treated as a predictive signal. First, we expand the existing EEGEyeNet corpus [13] by collecting and releasing new data for *naturalistic paradigms*, specifically watching movies and reading sentences, using *two different hardware* setups. Combined with existing and publicly available datasets, this multi-paradigm, multi-hardware dataset provides the scale needed to train models on the complex ocular patterns of various behaviors. Second, we introduce a unified, end-to-end model that reconstructs complete scanpaths directly from EEG in a single forward pass. While prior work has shown that specific properties such as saccade direction or fixation position can be decoded in isolation [14, 13], our approach uses a unified loss to jointly predict ocular event types, temporal boundaries, and spatial displacements (see Figure 1). We further demonstrate that this model can be effectively adapted to unseen users or hardware using a "minimal calibration (fine-tuning)" of only 5 minutes of data, leading to a further improved accuracy. By combining this expanded data resource with a unified architecture, we provide a practical approach to leveraging gaze information in legacy datasets and clinical settings where dedicated eye-tracking is not feasible.

2 Methods

Given a continuous EEG window, we predict an ordered set consisting of fixations, saccades, and blinks and their spatial and temporal features. Each event is represented by its class, onset and offset times, and event-specific spatial parameters: absolute gaze position for fixations (XY coordinates), and displacement endpoints for saccades and blinks (relative change represented by the angle and amplitude of eye movements from previous state). The task therefore combines temporal event segmentation with spatial regression from EEG.

Dataset We build on EEGEyeNet [13], which contains high-density EEG and synchronized eye-tracking recordings from Dots, ANTIsaccade (ANTI), and Visual Symbol Search (VSS). We extend it with additional Dots recordings and two new naturalistic paradigms: Natural Reading (NR) and Movie Watching (MW). The resulting dataset spans five paradigms and two EEG hardware configurations, Geodesic Hydrocel and ANT Neuro, allowing us to test generalization across both paradigms and acquisition systems. To enable this cross-hardware evaluation despite differing electrode layouts, we

Table 1: Dataset statistics per evaluation setting and paradigm. We report the number of available samples (#Fixations, #Saccades, #Blinks) for each paradigm. Full paradigm descriptions are provided in Appendix C.

Dataset	Dots	VSS	ANTI	NR	MW	Total
# Subjects	72	219	226	34	61	324
# Fixations	159 093	192 000	445 227	62 246	40 927	899 493
# Saccades	133 616	185 993	371 717	58 142	35 491	784 959
# Blinks	24 049	5 217	72 646	4 073	5 396	111 381

spatially map the channels between the two systems based on the minimal physical distance on the scalp (see Appendix D for details on the cross-hardware mapping of electrodes).

Unlike the original benchmark, which primarily evaluates controlled gaze decoding, we evaluate all spontaneously occurring fixations, saccades, and blinks across paradigms and hardware setups. Recording details are provided in Appendix C. Dataset statistics are summarized in Table 1, and event-distribution analyses are shown in Figures 9, 10, 11, 12, 13.

Metrics We evaluate predictions using three metrics: macro F1 over fixation, saccade, and blink classes; visual distance (VD) for event-level spatial error; and Dynamic Time Warping (DTW) for full-sequence scanpath alignment. The table-specific interpretation of these metrics is given with the benchmark results, and formal definitions, including the visual-angle conversion and DTW distance, are provided in Appendix D.3.

Data Splits and Training Setups We use an across-subject 80/20 train-test split, ensuring that no participant appears in both training and test sets. Hyperparameter choices and model selection are performed exclusively on the training set (see Appendix B for details), ensuring that the test set remains fully held out and is used only for final evaluation. For participants recorded in multiple paradigms, all sessions are assigned to the same split to prevent cross-paradigm leakage. All experiments use the same fixed test set, which contains data from all paradigms and hardware setups. We evaluate three training regimes. **Dots-only** trains on the Dots paradigm and tests both within-paradigm and cross-paradigm generalization. **ALL** trains on all available paradigms: Dots, VSS, ANTI, NR, and MW. **Calibration** starts from a pretrained model and fine-tunes it using 5–25 minutes of target-user data, measuring adaptation to new subjects and to the ANT Neuro hardware. Calibration details are provided in Appendix G.

3 Model

In this section, we propose a unified model for end-to-end scanpath reconstruction. Because existing methods primarily focus on isolated gaze-decoding sub-tasks (decoding of saccade or fixation features only), we evaluate our approach against the best models in each subtask as well as a constructed two-stage baseline (Figure 2) for the decoding of the full scanpath.

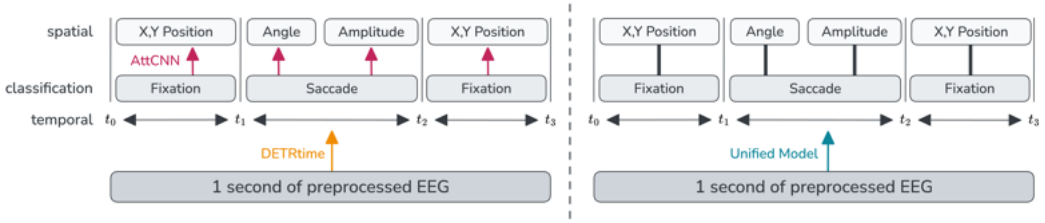


Figure 2: The standard two-stage pipeline (left) first performs temporal segmentation (↑) and then applies spatial decoding independently (↑). Our unified scanpath decoder (right) jointly predicts event type, timing, and spatial features in a single forward pass (↑).

3.1 Baselines

For baseline comparisons, we use state-of-the-art models from the EEGEyeNet benchmark suite [13]. We select *DETRtime* for temporal event segmentation into fixations, saccades, and blinks [15], and *Attention-CNN* for spatial decoding of eye-movement features such as fixation position, saccade amplitude, and angle [16]. While these models are well-suited for their respective single tasks (spatial feature prediction or temporal segmentation), they do not provide a unified framework for full scanpath decoding. Because these models solve separate single-tasks, we combine them into a two-stage scanpath baseline, illustrated in Figure 2. *DETRtime* first segments the continuous EEG into event windows, after which *Attention-CNN* predicts spatial features for each detected event. The predicted events are then concatenated in temporal order to reconstruct the scanpath. This provides a modular baseline, but temporal and spatial decoding are optimized independently.

3.2 Unified Model

We propose a unified model that adapts DETRtime from temporal segmentation to full scanpath decoding. The model processes raw EEG with a one-dimensional convolutional backbone, adds sinusoidal positional encoding, and uses a transformer encoder-decoder with a fixed set of learned event queries, as shown in Figure 3. Each query predicts one ocular event, allowing a fixed-size output to represent a variable-length scanpath through an additional dummy "no-event" class.



Figure 3: Unified Model Architecture consisting of convolutional backbone, detection transformer and MLP heads to decode all temporal and spatial features in one forward pass.

Architecture and event predictions. The transformer uses six encoder layers, six decoder layers, eight attention heads, and hidden dimension 128. Each decoder query attends to the encoded EEG segment and outputs three quantities: an *event type* distribution over fixation, saccade, blink, and dummy classes; a *temporal extent* given by normalized onset and offset times; and *spatial features*. Fixations are represented by absolute gaze position, while saccades and blinks are represented as displacements from the preceding fixation. This representation preserves scanpath continuity and allows post-blink drift to be modeled. Spatial features are predicted by dedicated MLP heads.

Training objective. The training objective is a sum of three per-pair terms, applied after the matching step described below,

$$\mathcal{L}_{\text{unified}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{temporal}} + \mathcal{L}_{\text{spatial}}.$$

Classification loss ($\mathcal{L}_{\text{CE}}$). A standard cross-entropy over the four classes (fixation, saccade, blink, dummy) supervises the event-type head.

Temporal loss ($\mathcal{L}_{\text{temporal}}$). For a predicted boundary interval $\hat{b}$ and its target b, we combine an ℓ_1 distance with the generalised IoU,

$$\mathcal{L}_{\text{temporal}} = \|b - \hat{b}\|_1 + (1 - \text{GIoU}), \qquad \text{GIoU}(b, \hat{b}) = \text{IoU}(b, \hat{b}) - \frac{|b_{\text{enc}}| - \text{Union}(b, \hat{b})}{|b_{\text{enc}}|},$$

where b_{enc} is the smallest interval enclosing b and $\hat{b}$. Unlike standard IoU, GIoU yields an informative gradient even when the two intervals do not overlap.

Spatial loss ($\mathcal{L}_{\text{spatial}}$). Every event is scored against an endpoint $\mathbf{p} = (p^x, p^y)$ on the screen: the gaze point for fixations and, for saccades and blinks, the displacement endpoint anchored at the target start $\mathbf{s}$. The model emits this point as a polar pair $(\hat{r}, \hat{\theta})$, expressed relative to the screen centre for fixations and to $\mathbf{s}$ otherwise, and is scored by a screen-normalised Euclidean distance,

$$\hat{\mathbf{p}} = \begin{cases} \hat{r}\,(\cos\hat{\theta},\ \sin\hat{\theta}) & \text{(fixation)}, \\ \mathbf{s} + \hat{r}\,(\cos\hat{\theta},\ \sin\hat{\theta}) & \text{(saccade/blink)}, \end{cases} \qquad \mathcal{L}_{\text{spatial}}(\mathbf{p}, \hat{\mathbf{p}}) = \sqrt{\left(\frac{\hat{p}^x - p^x}{W}\right)^2 + \left(\frac{\hat{p}^y - p^y}{H}\right)^2}$$

with screen width W and height H. The normalization makes the loss dimensionless and display-independent.

Three design axes determine this form: the *coordinate system* ((x, y) vs. (r, θ)), the *loss formulation* (unified Euclidean vs. disaggregated ℓ_1), and the *granularity of spatial supervision* (all event types vs. fixations only, with saccades and blinks interpolated post-hoc). All eight combinations are evaluated in the ablation study (Section 4.4), where the polar, unified-Euclidean, all-heads variant used here achieves the best average DT. Further details are provided in Appendix E.

Optimization and inference. We train with Adam using a learning rate of 5×10^{-5} and batch size 32. During training, dynamic sampling balances event classes and paradigms. At inference, dummy predictions are discarded, remaining events are ordered by time, overlaps are resolved by confidence, and uncovered intervals are assigned to the previous event. The resulting event sequence is used as the predicted scanpath.

4 Experiments

We benchmark our unified model against single-task baselines and compare inference times. We then evaluate fine-tuning for calibration, analyze prediction errors regarding saccade amplitude, angle, and the 2D spatial distribution, and finally perform an ablation study of our modeling choices.

4.1 Benchmark

Table 2 reports results under two training conditions: limited training on the Dots paradigm only, and full training across all paradigms (ALL).

Table 2: Evaluation results comparing single-task (S) and unified (U) models across six evaluation datasets. Rows are grouped by training dataset (Dots or ALL) and metric. We report F1 score for event segmentation, Visual Distance (VD) in degrees for saccade and fixation endpoint errors, and Dynamic Time Warping (DTW) distance in degrees for full scanpath similarity. Higher F1 ↑ and lower VD/DTW ↓ indicate better performance.

Dataset→ ↓			**Evaluation**					
Training	**Metric**	**Model**	**DOTS**	**NR**	**MW**	**VSS**	**ANTI**	**ALL**
Dots	Segmentation (F1) ↑	S	0.78	**0.78**	**0.77**	**0.69**	0.72	**0.75**
		U	**0.79**	0.77	**0.77**	0.67	**0.73**	**0.75**
	Saccade (VD) ↓	S	3.49	3.21	3.54	6.29	4.59	4.22
		U	**3.00**	**2.78**	**2.69**	**5.82**	**3.67**	**3.59**
	Fixation (VD) ↓	S	**4.62**	5.34	4.31	**6.86**	4.11	5.05
		U	4.69	**5.22**	**4.05**	7.45	**3.51**	**4.98**
	Scanpath (DTW) ↓	S	9.47	8.13	5.80	11.53	7.03	8.39
		U	**6.22**	**4.32**	**3.09**	**5.49**	**3.41**	**4.51**
ALL	Segmentation (F1) ↑	S	0.81	0.84	0.80	0.70	**0.78**	0.79
		U	**0.82**	**0.86**	**0.83**	**0.72**	**0.78**	**0.80**
	Saccade (VD) ↓	S	2.84	3.04	3.35	5.79	4.34	3.87
		U	**2.45**	**1.24**	**1.81**	**1.84**	**1.90**	**1.85**
	Fixation (VD) ↓	S	5.99	6.09	3.53	7.26	3.14	5.20
		U	**4.89**	**3.96**	**3.36**	**5.67**	**2.20**	**4.02**
	Scanpath (DTW) ↓	S	9.21	8.15	5.75	11.24	5.08	7.89
		U	**5.43**	**5.19**	**2.30**	**7.51**	**2.99**	**4.68**

Note that the unified model jointly predicts event detection and spatial features, and therefore does not produce spatial outputs for events it fails to detect. To ensure a fair comparison for event-level spatial metrics (saccade and fixation metrics in the table), we evaluate both approaches only on the

same subset of correctly detected events. This ensures that both the unified model and the baselines are evaluated under identical conditions and under the exact same evaluation set for spatial decoding. In contrast, segmentation and scanpath-level metrics (e.g., DTW) are not subject to this restriction and are computed on the full set of evaluation set, capturing overall end-to-end performance. Further details are provided in Appendix A.

Unified model outperforms single-task pipelines. When restricted to Dots-only training, on average the unified decoder matches or surpasses the score of single task baselines (last column of the table marked by **ALL**), although performance is less consistent across paradigms (e.g VSS). Nonetheless, when training with ALL paradigms (the lower part of the table), unified model performs better than single task models on every metric and any condition, with the largest gains on naturalistic data: DTW drops from 5.75° to 2.30° on MW and from 5.08° to 2.99° on ANTI. These DTW improvements reflect coherent fixation–saccade sequences rather than isolated event-level wins, and confirm that jointly modelling segmentation and spatial decoding scales with paradigm diversity.

Generalization improves with multi-paradigm training. When comparing unified models trained on Dots-only versus those trained on all paradigms, we observe substantial gains in generalization. The largest improvements emerge in naturalistic conditions such as Natural Reading and Movie Watching. For example, saccade error in NR drops from 2.78° to 1.24°, and in MW from 2.69° to 1.81°. Fixation decoding also improves significantly, with error in NR decreasing from 5.22° to 3.96°, and in ANTI from 3.51° to 2.20°. These gains suggest that training the unified model on diverse paradigms yields more robust and generalizable EEG-to-gaze representations.

Overall performance. Despite the inherent challenges of EEG-based gaze prediction (such as low spatial resolution, noise, and variability across subjects and tasks) our unified model achieves competitive results across multiple metrics. When trained on the full set of paradigms, it reaches a mean saccade error of 1.85°, a fixation error of 4.02°, and a segmentation F1 score of 0.80. The average DTW scanpath distance is 4.68°, outperforming all single-task baselines. While fixation accuracy remains moderate, the ability to decode saccades with sub-2° error highlights the model's sensitivity to relative gaze shifts—an encouraging result for non-invasive, EEG-only gaze tracking.

Efficiency. Our unified decoder predicts event type, timing and spatial coordinates in a single forward pass at almost no cost over a segmentation-only DETRtime baseline: 8.1 M vs. 8.0 M parameters and 65.25 ms vs. 65.93 ms inference per 1-second EEG window on an NVIDIA Tesla T4. The corresponding single-task pipeline—DETRtime followed by an AttentionCNN (6.0 M, 28.12 ms) applied to each detected segment—requires sequential forward passes and totals 94.05 ms, i.e. $\sim 1.4\times$ the unified-model latency, while delivering worse accuracy across all metrics (Table 2).

4.2 Calibration via Fine-Tuning

The previous section established that the unified model already generalizes well to subjects unseen during training. We now ask a complementary, practical question: *how much of the remaining gap to a subject-specific model can be closed with a short per-user calibration step?* We study this in two progressively harder regimes: (i) *cross-user* calibration, where the unified model is fine-tuned on DOTS data from a held-out Hardware 1 subject; and (ii) *cross-hardware* calibration, where the model pretrained on Hardware 1 is fine-tuned on subjects recorded with different EEG hardware (ANT Neuro; Hardware 2). We vary the amount of per-user calibration data from 5 to 25 minutes of DOTS recordings in 5 min increments, with one of the user's sessions held out for evaluation.

Cross-User Calibration. Starting from a competitive baseline (F_1=0.80, Fixation VD 4.91°, Saccade VD 2.65°, DTW 4.84°) across N=12 held-out Hardware 1 users, just 5 minutes of per-user DOTS data ($\approx$300 samples) reduces Fixation VD by 11.8%, Saccade VD by 13.3% and Scanpath DTW by 12.7%. Most of the achievable gain can be observed within 10 minutes of training data, at which point the errors have dropped by 17.5%, 23.0% and 18.6% respectively. This is close to what 25 minutes of training data eventually delivers (20.1%, 23.5% and 20.6%), bringing the model to 3.92° Fixation VD, 2.03° Saccade VD and 3.84° DTW at F_1=0.85 (a $+0.05$ gain over the Baseline).

Cross-Hardware Calibration. The cross-hardware regime is substantially harder: the unseen acquisition hardware (ANT Neuro) introduces signal statistics absent from the training distribution,

and on the $N{=}8$ held-out Hardware 2 users the Baseline suffers a sizeable shift ($F_1{=}0.64$, Fixation VD $8.15°$, Saccade VD $7.27°$, DTW $8.17°$). Here, calibration proves the strength and flexibility of our model. After only 5 min of DOTS data per user, Fixation VD drops by 31.5%, Saccade VD by 50.4% and DTW by 32.9%, while F_1 rises by $+0.124$. Gains continue up to $\sim$15 min, after which the curves plateau. At 25 min the model reaches $F_1{=}0.85$, $4.48°$ Fixation VD, $2.28°$ Saccade VD and $4.41°$ DTW—reductions of 45.1%, 68.7% and 46.0% over the Baseline, and an F_1 gain of $+0.21$. These calibration results show that the performance loss caused by subject- or hardware-specific variability is largely recoverable through a brief per-user calibration. A practical system can be bootstrapped from the unified model with as little as $5{-}10$ minutes of subject-specific data, and extended to $\sim$25 minutes when deployment introduces unseen hardware characteristics.

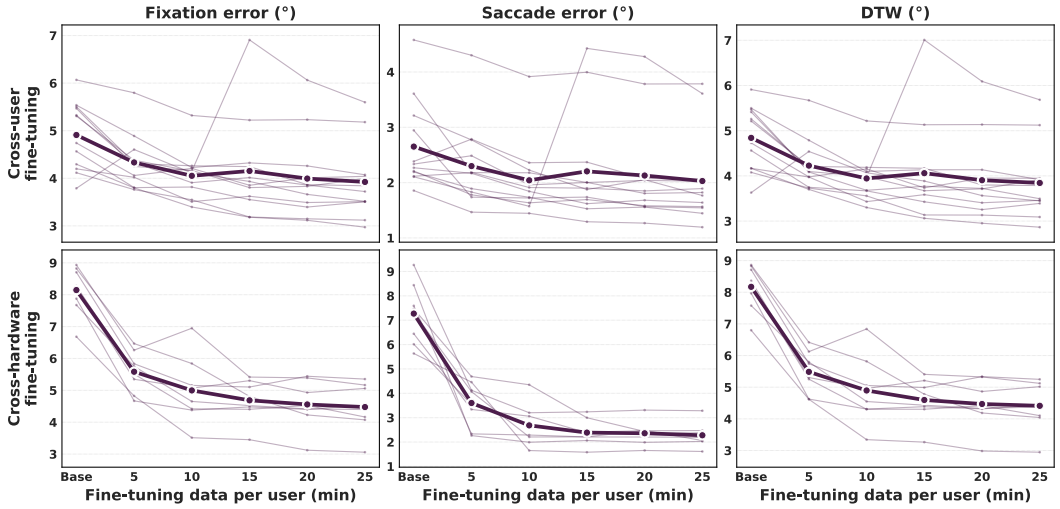
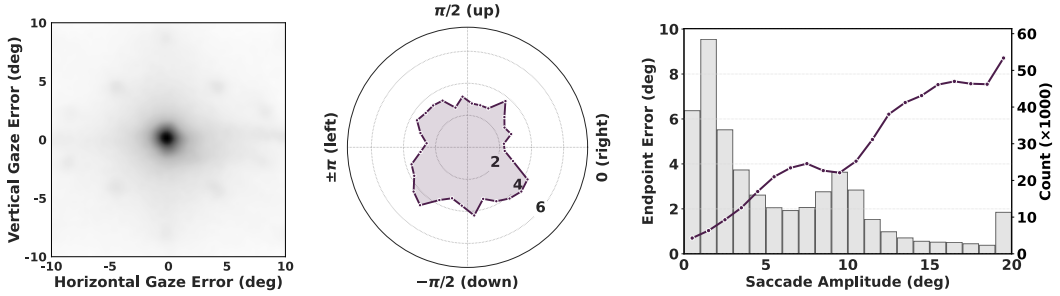


Figure 4: Calibration curves. Fine-tuning trajectories for the cross-user (top) and cross-hardware (bottom) settings. Light lines show individual subjects; the bold line is the across-user average. In both settings, most of the improvement is gained within the first $5{-}10$ minutes of calibration data.

4.3 Error Analysis

In the following, we analyze how decoding performance varies with respect to three key spatial factors. We focus on three complementary aspects of spatial decoding performance: (i) the relationship between saccade amplitude and prediction error, (ii) directional biases in angular decoding, and (iii) the spatial distribution of decoded gaze positions relative to the ground-truth origin $(0, 0)$. These results are aggregated across all paradigms, with a detailed breakdown provided in the Appendix.



(a) Spatial error (b) Angular error (c) Amplitude vs error

Figure 5: Error analysis across saccade amplitude, angle, and 2D spatial error distribution, aggregated across all paradigms.

Impact of Saccade Amplitude. We first analyze how decoding error varies with saccade amplitude. As shown in Figure 5 (right), the error increases approximately monotonically with movement magnitude. Small-amplitude saccades ($< 2^\circ$) exhibit consistently low error, while larger movements introduce higher uncertainty. This trend is consistent across paradigms and hardware setups, indicating that spatial decoding difficulty scales with the magnitude of neural displacement (see Appendix F for a detailed breakdown for each paradigm).

Impact of Saccade Direction. We next examine directional biases in decoding performance. Figure 5 (middle) shows the angular error distribution in polar coordinates. Overall, the model exhibits stable performance across directions, with mild reductions in error along horizontal axes compared to oblique and vertical movements. This suggests that horizontal eye movements are more robustly encoded in EEG signals (see Appendix F for a detailed breakdown for each paradigm).

Spatial Distribution of Decoding Error. Finally, we evaluate the global spatial structure of prediction errors. As shown in Figure 5 (left), decoded gaze positions form a tightly centered distribution around the ground-truth origin $(0, 0)$. While the majority of predictions remain close to the center, we observe slightly increased dispersion along the vertical axis for more complex viewing conditions, suggesting asymmetric sensitivity in vertical gaze decoding (see Appendix F for a detailed breakdown for each paradigm).

4.4 Ablation Study

To better understand the contributions of different architectural and training choices, we conduct a series of eight ablation experiments using the unified decoder. These ablations test the three modelling axes introduced in the loss definition section: the *coordinate system* in which spatial attributes are predicted, the *loss formulation* applied to those attributes, and the *prediction head configuration* (i.e. the granularity of spatial supervision across event types). For coordinates, we compare the polar parametrisation (r, θ) against a direct Cartesian parametrisation (x, y), related by $(x, y) = r\,(\cos\theta, \sin\theta)$. Both encode the same geometry but induce different loss landscapes: Cartesian supervision is piecewise linear in the network outputs, whereas polar supervision routes gradients through the nonlinear $(\cos, \sin)$ mapping. For the loss formulation, we contrast the unified Euclidean objective $\mathcal{L}_{\text{spatial}}$ in Eq. (3.2) with a disaggregated ℓ_1 variant that scores each spatial component independently,

$$\mathcal{L}^{\ell_1}_{\text{spatial}} = \frac{1}{|\mathcal{E}^{\text{tgt}}|} \sum_{e_j \in \mathcal{E}^{\text{tgt}}} w_{\ell_j} \left(\frac{|\hat{u}_{j,1} - u_{j,1}|}{N_1} + \frac{|\hat{u}_{j,2} - u_{j,2}|}{N_2} \right), \tag{1}$$

where $(\hat{u}_1, \hat{u}_2)$ and (u_1, u_2) are predicted and target spatial components in the chosen coordinate system: (x, y) (or $(\Delta x, \Delta y)$) in the Cartesian variant with $(N_1, N_2) = (W, H)$, and (r, θ) in the polar variant with N_1 the screen diagonal and $N_2 = 1$. For the angular component we use the circular ℓ_1 distance $d_\theta(\hat{\theta}, \theta) = |\text{atan2}(\sin(\hat{\theta} - \theta), \cos(\hat{\theta} - \theta))|$ so that $\theta = -\pi$ and $\theta = \pi$ receive zero penalty. This disaggregated form removes the coupling between the two spatial dimensions but also the rotation invariance of the endpoint error. Finally, for head configuration, we compare the *all-heads* setting, in which $\mathcal{L}_{\text{spatial}}$ is applied to every event class—fixations (absolute position), saccades and blinks (relative displacement)—against a *fixation-only* setting, where saccades and blinks contribute to the temporal/classification losses but not to $\mathcal{L}_{\text{spatial}}$, and their trajectories are reconstructed at decoding time by linear interpolation between consecutive fixation anchors.

Configuration			Paradigm					
Coordinate system	Loss	Heads	AVERAGE	DOTS	NR	MW	VSS	ANTI
Cartesian	Euclidean	all	4.75	4.36	4.81	4.04	7.11	3.41
Cartesian	Euclidean	fixation only	4.72	4.55	5.08	3.70	7.18	3.11
Polar	**Euclidean**	**all**	**4.51**	**6.22**	**4.32**	**3.09**	**5.49**	**3.41**
Polar	Euclidean	fixation only	4.89	6.81	4.77	3.40	5.83	3.65
Cartesian	L1	all	5.72	5.15	5.70	4.76	8.36	4.62
Cartesian	L1	fixation only	5.59	5.54	5.54	4.93	7.62	4.32
Polar	L1	all	4.90	7.24	4.48	3.31	5.73	3.76
Polar	L1	fixation only	7.80	8.90	7.77	6.41	9.28	6.66

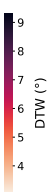


Figure 6: DTW results across different model configurations.

Our results (Figure 6) show that the best performance across datasets is achieved when using all event heads, polar coordinates, and the Euclidean loss. Notably, although it is possible to reconstruct full scanpaths using only fixation predictions (by interpolating saccades between successive fixations) this variant underperforms, especially in naturalistic tasks. The Euclidean loss consistently outperforms the disaggregated ℓ_1 loss, highlighting the advantage of a unified spatial objective that couples the two spatial dimensions through a single rotation- and scale-consistent endpoint error and avoids task-specific branching in the loss function. We hypothesise that this simplification leads to a more stable and coherent learning process. One exception is observed in the Dots dataset, where the Cartesian setup performs competitively as well.

5 Related Work

Classical eye-event detection and gaze decoding from other modalities. Most ocular event labels in EEG datasets still come from infrared video trackers paired with rule-based dispersion/velocity/acceleration detectors [17, 18]. These are millidegree-precise but fail without an infrared camera — a constraint that excludes many archival EEG datasets, infants, and low-vision cohorts. Hardware-free alternatives include fMRI voxel decoding, where retinotopic patterns predict coarse gaze [19] and *DeepMReye* reconstructs 2D point-of-regard from MR images of the eyes [20], but temporal resolution is capped by the hemodynamic response and the setup is restricted to MRI facilities. Classical EEG/EOG pipelines instead treat ocular artefacts as the signal, using probabilistic methods [21] to detect blinks, saccades and fixations directly from EOG, with Dimigen et al. [4] demonstrating millisecond-accurate co-registration during natural reading. Such pipelines rely on per-subject thresholds and rarely produce full scanpaths or sub-degree spatial estimates.

Deep learning for EEG-based gaze inference. End-to-end deep networks on raw EEG have largely targeted a single gaze attribute on tightly controlled paradigms. Kastrati et al. classified saccade direction with a shallow CNN [14]; the *EEGEyeNet* benchmark [13] and its follow-up *DETRtime* [15] introduced an Attention-CNN for spatial regression and a transformer for temporal segmentation — each solving its task in isolation. More recent work extends this line architecturally (spatial-filtering CNNs [22], MobileViT distillations for on-device inference [23]) and through training-time supervision (paired eye-tracker labels used only during training, reaching $\sim 0.9°$ median error on grid-like data [24]). Across this family, models almost always decode a single attribute, train and evaluate on a single paradigm, and rarely report cross-paradigm or cross-subject generalisation — the gaps that motivate this work.

Unified multi-task architectures. In computer vision, single networks that jointly solve several prediction tasks — Mask R-CNN [25], DETR [26], Mask2Former [27], SAM [28] — consistently show that diverse data and multi-task heads yield models that generalise widely. Our scanpath decoder transposes this philosophy to EEG, jointly producing event labels, onset/offset timing, and spatial displacements from neural time-series in a single forward pass.

6 Conclusion

We demonstrate that eye-movement-related activity in EEG (traditionally treated as an artifact) can be used to reconstruct scanpaths end-to-end. By introducing a unified model that jointly predicts event type, timing, and spatial structure, we show that these signals can be decoded reliably at the single-trial level, outperforming modular pipelines that treat these components independently. Our large-scale evaluation across five diverse paradigms and two distinct hardware setups shows that the unified decoder consistently improves over strong single-task baselines. Performance further improves with multi-paradigm training or finetuning, highlighting the importance of dataset diversity for learning generalizable brain-to-gaze mappings. The model achieves sub-2° saccade error, 4° fixation error, and maintains robust segmentation performance (F1 score at 0.80) across unseen participants — even without per-subject calibration. Questions remain about the neural sources of these decoded signals and whether performance can be further improved with higher-density EEG arrays, or adapted architectures. Future work should explore the interpretability of learned features, and assess performance in truly mobile settings. As EEG hardware continues to miniaturize, this approach could enable pervasive gaze tracking in clinical assessment, brain-computer interfaces, and augmented reality applications where cameras are impractical or to complement them in hybrid systems.

References

[1] Keith Rayner. Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, 62(8):1457–1506, 2009.

[2] Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost van de Weijer. *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford University Press, Oxford, UK, 2011.

[3] Beatriz Luna, Katerina Velanova, and Charles F Geier. Development of eye-movement control. *Brain and cognition*, 68(3):293–308, 2008.

[4] Olaf Dimigen, Werner Sommer, Annette Hohlfeld, Arthur M. Jacobs, and Reinhold Kliegl. Coregistration of eye movements and eeg in natural reading: analyses and review. *Journal of Experimental Psychology: General*, 140(4):552–572, 2011.

[5] Michael Plöchl, José P. Ossandón, and Peter König. Combining eeg and eye tracking: identification, characterization, and correction of eye movement artifacts in electroencephalographic data. *Frontiers in Human Neuroscience*, Volume 6 - 2012, 2012.

[6] Matthias S. Treder and Benjamin Blankertz. (C)overt attention and visual speller design in an ERP-based brain-computer interface. *Behavioural and Brain Functions*, 6(1):28, 2010.

[7] Ard Kastrati, R. Michael Winters, Nemanja Djuric, Ivan J. Tashev, and Yu-Te Wang. Automated feature engineering for single-trial eeg and eye-tracking classification in predictive text interfaces. In *2025 13th International Conference on Brain-Computer Interface (BCI)*, pages 1–7, 2025.

[8] Johannes Zagermann, Ulrike Pfeil, and Harald Reiterer. Measuring cognitive load using eye tracking technology in visual computing. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, BELIV '16, page 78–85, New York, NY, USA, 2016. Association for Computing Machinery.

[9] Wolfgang Fuhl, Marc Tonsen, Andreas Bulling, and Enkelejda Kasneci. Pupil detection for head-mounted eye tracking in the wild: an evaluation of the state of the art. *Machine Vision and Applications*, 27(8):1275–1288, 2016.

[10] Roy S. Hessels and Ignace T. C. Hooge. Eye tracking in developmental cognitive neuroscience—the good, the bad and the ugly. *Developmental Cognitive Neuroscience*, 40:100710, 2019.

[11] Kenneth Holmqvist, Marcus Nyström, and Fiona Mulvey. Eye-tracker data quality: what it is and how to measure it. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '12)*, pages 45–52, 2012.

[12] Olaf Dimigen and Benedikt Ehinger. Regression-based analysis of combined eeg and eye-tracking data: Theory and applications. *Journal of Vision*, 21:3, 01 2021.

[13] Ard Kastrati, Martyna Beata Plomecka, Damián Pascual, Lukas Wolf, Victor Gillioz, Roger Wattenhofer, and Nicolas Langer. Eegeyenet: a simultaneous electroencephalography and eye-tracking dataset and benchmark for eye movement prediction. *CoRR*, abs/2111.05100, 2021.

[14] Ard Kastrati, Martyna Beata Plomecka, Roger Wattenhofer, and Nicolas Langer. Using deep learning to classify saccade direction from brain activity. In *ACM Symposium on Eye Tracking Research and Applications (ETRA '21)*. ACM, 2021.

[15] Lukas Wolf, Ard Kastrati, Martyna Beata Płomecka, J. Li, Dustin Klebe, Alexander Veicht, Roger Wattenhofer, and Nicolas Langer. A deep learning approach for the segmentation of electroencephalography data in eye tracking applications. In *International Conference on Machine Learning*, 2022.

[16] Nina Weng, Martyna Plomecka, Manuel Kaufmann, Ard Kastrati, Roger Wattenhofer, and Nicolas Langer. An interpretable and attention-based method for gaze estimation using electroencephalography. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023: 26th International Conference, Vancouver, BC, Canada, October 8–12, 2023, Proceedings, Part II*, page 734–743, Berlin, Heidelberg, 2023. Springer-Verlag.

[17] Marcus Nyström and Kenneth Holmqvist. An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior Research Methods*, 42(1):188–204, 2010.

[18] Ralf Engbert and Reinhold Kliegl. Microsaccades uncover the orientation of covert attention. *Vision Research*, 43(9):1035–1045, 2003.

[19] Joon Son, Lei Ai, Ryeongkwon Lim, Ting Xu, Stan Colcombe, Alexander R Franco, Jessica Cloud, Stephen LaConte, Jacob Lisinski, Arno Klein, R. Cameron Craddock, and Michael Milham. Evaluating fmri-based estimation of eye gaze during naturalistic viewing. *Cerebral Cortex*, 30(3):1171–1184, Mar 2020.

[20] Markus Frey, Matthias Nau, and Christian F. Doeller. Magnetic resonance-based eye tracking using deep neural networks. *Nature Neuroscience*, 24(12):1772–1779, 2021.

[21] Miika Toivanen, Kati Pettersson, and Kristian Lukander. A probabilistic real-time algorithm for detecting blinks, saccades, and fixations from eog data. *Journal of Eye Movement Research*, 8(2):1–15, 2015.

[22] Wolfgang Fuhl, Susanne Zabel, Theresa Harbig, Julia Astrid Moldt, Teresa Festl Wiete, Anne Herrmann Werner, and Kay Nieselt. One step closer to eeg-based eye tracking. *arXiv preprint arXiv:2303.06039*, 2023.

[23] Teng Liang and Andrews Damoah. Eegmobile: Enhancing speed and accuracy in eeg-based gaze prediction with advanced mobile architectures. *arXiv preprint arXiv:2408.03449*, 2024. Accepted *HCI International 2024 – Late Breaking Work*.

[24] Rui Sun, Andy S. K. Cheng, Cynthia Chan, Janet H. W. Hsiao, Anthony J. Privitera, Jun Gao, Calvin H. Fong, Ren Ding, and Akaysha C. Tang. Tracking gaze position from eeg: Exploring the possibility of an eeg-based virtual eye-tracker. *Brain and Behavior*, 13(10):e3205, 2023.

[25] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[26] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[27] Bowen Cheng, Alexander Schwing, and Alexander Kirillov. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.

[28] Alexander Kirillov, Eric Mintun, Nikhila Ravi, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[29] Chrystalina Antoniades, Ulrich Ettinger, Bertrand Gaymard, Iain Gilchrist, Arni Kristjánsson, Christopher Kennard, R John Leigh, Imran Noorani, Pierre Pouget, Nikolaos Smyrnis, et al. An internationally standardised antisaccade protocol. *Vision research*, 84:1–5, 2013.

[30] David Wechsler and Habuku Kodama. *Wechsler intelligence scale for children*, volume 1. Psychological corporation New York, 1949.

[31] Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. Zuco 2.0: A dataset of physiological recordings during natural reading and annotation. *CoRR*, abs/1912.00903, 2019.

[32] Emily S Kappenman and Steven J Luck. The effects of electrode impedance on data quality and statistical significance in erp recordings. *Psychophysiology*, 47(5):888–904, 2010.

[33] Andreas Keil, Stefan Debener, Gabriele Gratton, Markus Junghöfer, Emily S Kappenman, Steven J Luck, Phan Luu, Gregory A Miller, and Cindy M Yee. Committee report: publication guidelines and recommendations for studies using electroencephalography and magnetoencephalography. *Psychophysiology*, 51(1):1–21, 2014.

A Benchmark

For completeness, and to allow comparison with prior work that reports results in millimeters, we include evaluation metrics in mm as well. In Table 2 (in degrees) as well as in Table 3, we benchmark our unified model against baselines across several tasks.

Fair Comparison for Fixations and Saccades. In the single-task baselines for saccade and fixation decoding, spatial predictions are computed on inputs that are already segmented using ground-truth event boundaries. This implies that every event is perfectly localized in time, and spatial decoding is performed under an idealized setting where no temporal uncertainty is present. In contrast, the unified model jointly performs temporal event detection and spatial feature regression. As a consequence, spatial predictions are only available for events that are successfully detected by the model; missed detections do not yield a corresponding spatial output. To enable a fair and meaningful comparison of spatial decoding performance, we evaluate both models only on the subset of events for which the unified model produces a prediction that overlaps with a ground-truth event. Spatial metrics (VD) are then computed on this shared set of matched events for both approaches. This design choice ensures that both models are evaluated under identical assumptions regarding temporal localization, thereby isolating spatial decoding quality from differences in event detection. However, it is worth mentioning that in absolute terms, this conditioning also introduces a selection effect, as the evaluation is restricted to successfully detected events and may therefore not fully reflect performance over the entire event distribution only in these subtask metrics. For DTW and F1 score, we do not have such constraints.

Table 3: Evaluation results comparing single-task (S) and unified (U) models across six evaluation datasets. Rows are grouped by training dataset (Dots or ALL) and metric. We report F1 score for event segmentation, Euclidean Distance (ED) in millimeters for saccade and fixation endpoint errors, and Dynamic Time Warping (DTW) distance in mm for full scanpath similarity. Higher F1 ↑ and lower ED/DTW ↓ indicate better performance.

Dataset→ ↓			**Evaluation**					
Training	**Metric**	**Model**	**DOTS**	**NR**	**MW**	**VSS**	**ANTI**	**ALL**
Dots	Segmentation (F1) ↑	S	0.78	**0.78**	**0.77**	**0.69**	0.72	**0.75**
		U	**0.79**	0.77	**0.77**	0.67	**0.73**	**0.75**
	Saccade (ED) ↓	S	41.93	38.59	42.59	76.91	55.36	60.94
		U	**36.26**	**33.51**	**32.11**	**70.39**	**44.00**	**46.00**
	Fixation (ED) ↓	S	**55.81**	64.20	51.37	**83.00**	49.32	58.70
		U	56.64	**62.63**	**48.33**	90.05	**42.07**	**53.89**
	Scanpath (DTW) ↓	S	114.01	97.49	69.28	139.08	84.17	95.85
		U	**74.65**	**51.54**	**36.77**	**65.72**	**40.68**	**51.00**
ALL	Segmentation (F1) ↑	S	0.81	0.84	0.80	0.70	**0.78**	0.80
		U	**0.82**	**0.86**	**0.83**	**0.72**	**0.78**	**0.80**
	Saccade (ED) ↓	S	34.10	36.56	40.12	70.49	52.34	56.07
		U	**29.59**	**14.96**	**21.57**	**22.41**	**22.82**	**23.38**
	Fixation (ED) ↓	S	72.26	73.03	42.03	87.88	37.63	57.43
		U	**59.06**	**47.57**	**40.10**	**68.66**	**26.38**	**41.39**
	Scanpath (DTW) ↓	S	110.86	97.56	68.70	135.59	60.78	80.69
		U	**65.11**	**61.94**	**27.41**	**90.07**	**35.78**	**48.05**

We also offer qualitative assessment of the performance of our Unified Scanpath Decoder. We provide example scanpath reconstructions and compare them to pipelines composed of single-task models. Figure 7 presents four representative sequences covering diverse scanpath patterns. The following examples include simple fixation sequences, long saccades, and complex event transitions.

These examples illustrate how the Unified Decoder achieves spatially precise and temporally coherent reconstructions, effectively capturing both the event segmentation and spatial trajectories within a

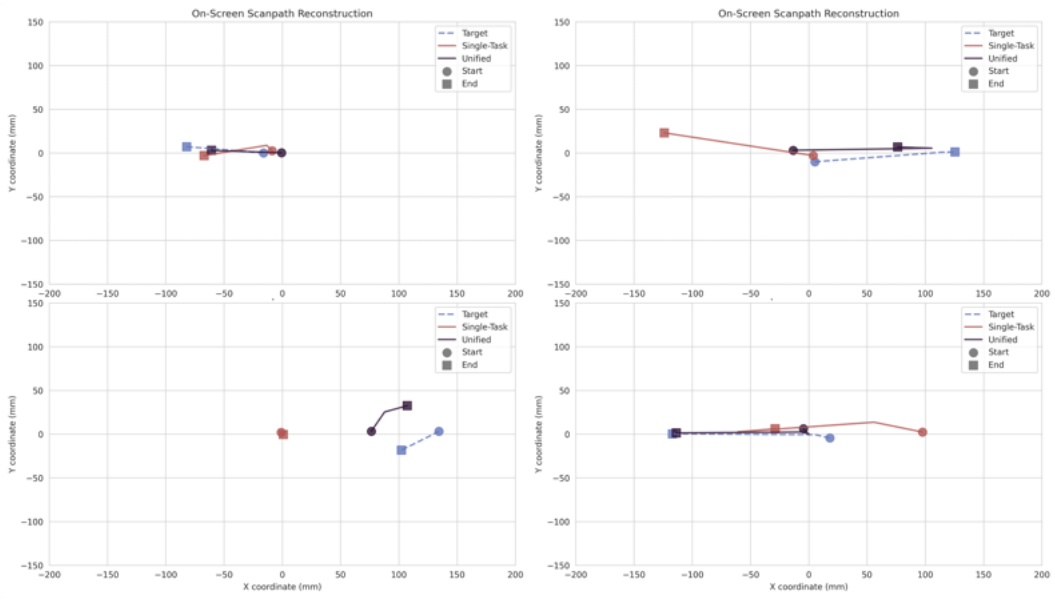


Figure 7: Examples of scanpath reconstructions from the Unified Decoder versus Single-Task pipelines. **Top-left**: Both pipelines perform comparably, but the Unified model better estimates the saccade angle. **Top-right**: The Single-Task pipeline fails to correctly decode the angle of a long saccade, while the Unified model accurately reconstructs it. **Bottom-left**: The Single-Task pipeline struggles to decode gaze positions far from the center, whereas the Unified model reconstructs a good scanpath, slightly spatially shift. **Bottom-right**: The Single-Task pipeline, unlike the Unified model, misaligns a long saccade in the middle of the scanpath, resulting in a noticeable shift.

single model, whereas single-task pipelines frequently suffer from segmentation errors and spatial drifts due to their decoupled nature.

B Hyperparameter Configurations and Training Setup

All experiments were conducted on high-performance GPUs, including NVIDIA H100 (80GB HBM3), A100-SXM4 (80GB), and Tesla V100-SXM2 (32GB). To optimize data loading and training throughput, we used up to 700GB of RAM to keep the dataset in memory during training. Each training run required approximately 1-6 days depending on the model and dataset size. We estimate the total compute for the experiments reported in the paper to be approximately 60 GPU-days. Significant additional computation was used for preliminary experiments not included in the final paper. For the hyperparameter choices we used the

Hyperparameter Description	**DETRtime**	**Unified Model**	**AttentionCNN**
Optimizer used for training	Adam	Adam	Adam
Learning rate	5×10^{-5}	5×10^{-5}	5×10^{-5}
Weight decay for regularization	10^{-2}	10^{-2}	10^{-2}
Type of backbone used	InceptionTime	InceptionTime	-
# filters in convolutional layers	64	64	-
Kernel size for convolutions	16	16	-
Residual connections	Yes	Yes	Yes
Depth of the backbone	4	4	-
# channels in the backbone output	16	16	-
# layers in the backbone	12	12	-
Total # Layers	-	-	12
Transformer hidden dimension	128	128	-
# object queries	20	20	-
# attention heads	8	8	-
# encoder layers in transformer	6	6	-
# decoder layers in transformer	6	6	-
Dimensionality of feedforward layers	2048	2048	-
Dropout rate	10%	10%	-
Pre-normalization in transformer	No	No	-
Target classes	3	3	3
Training epochs	30	30	30
Batch size	32	32	32
Biased sampling	Yes	Yes	Yes
Weight for class matching matrix	1	5	-
Weight for bounding box matching matrix	10	5	-
Weight for GIoU matching matrix	2	2	-
Cost for scanpath-feature matching matrix	-	2	-
Classification loss weight	1	4.0	-
Bounding box loss weight	4.0	4.0	-
GIoU loss weight	2.0	2.0	-
Loss weight for scanpath features	-	1.0	-

Table 4: Comparison of DETRtime, Unified Model and AttentionCNN hyperparameters

We use the default hyperparameters given by AttentionCNN and DETRtime. The DETRtime and Unified Model architectures share a common backbone (InceptionTime) and transformer-based design, and we use the same hyperparameters for these building blocks. Both process sequences with 64 convolutional filters, a kernel size of 16, and 12 backbone layers. Their transformer components are identical in structure. They feature 6 encoder and decoder layers, 8 attention heads, and a hidden dimension of 128. We borrow the exact same hyperparameters including learning rate, number of epochs, batch size and so on. However, the Unified Model differs slightly in its loss configuration, giving more balanced importance to the individual sub-matrices in the matching phase. The hyperparameters for the loss weights are chosen based on small scale experiment in the training set (without using the final evaluation set).

C Experimental Setup and Dataset

The study was conducted in a quiet, sound-insulated chamber with minimal lighting. Participants sat 68 cm from a 24-inch ASUS ROG Swift PG248Q display (viewable area: 531 × 299 mm). The monitor was operated at a resolution of 800 × 600 pixels—equivalent to 400 × 298.9 mm on screen—and a refresh rate of 100 Hz. Each participant completed the task individually while researchers monitored their performance from an adjacent control room. The protocol was approved by the local ethics committee.

C.1 Data Acquisition

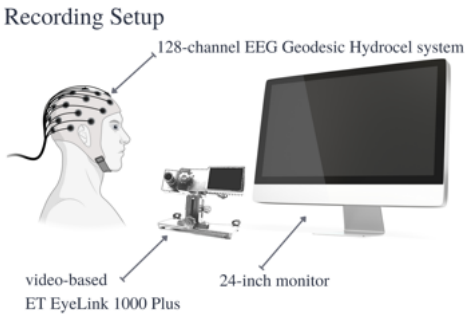


Figure 8: The illustration of the recording setup

Participants Data were obtained from 463 neurologically healthy adults (126 women) aged 20–80 years, who took part in four experimental paradigms (see Section C.2). Before the session, each volunteer provided their written informed consent covering both participation and future use of their data. Compensation was set at the local-currency equivalent of USD 25 per hour.

Eye-Tracking Acquisition Gaze data were collected with an infrared, video-based eye tracker (EyeLink 1000 Plus, SR Research) sampling at 500 Hz with an instrumental spatial resolution of 0.01 degree. A nine-point calibration was performed at the start of the session and revalidated before every trial block. The participants were instructed to fix the target until it vanished; calibration was accepted only if the mean calibration-validation error was below 1 degree of visual angle, otherwise the procedure was repeated.

EEG Acquisition High-density EEG was recorded at 500 Hz (band-pass 0.1–100 Hz) using a 128-channel Geodesic HydroCel system, with Cz as the recording reference. Electrode impedance was maintained below 40 kΩ and re-checked roughly every 30 minutes, lowering it when necessary.

C.2 Stimuli & Experimental Design

Dots For this paradigm, we used data from the EEGEyeNet dataset [13], where participants sequentially fixated dots displayed for 1.5–1.8 s at 25 distinct screen locations. The targets covered the four corners and the central region of the display, closely mirroring the spatial layout typically used for eye-tracker calibration. To enrich spatial sampling, the central dot was shown three times, giving 27 fixations per block. Each participant completed five pseudo-randomised blocks (45 s per block) to prevent anticipation of subsequent targets, and this entire sequence was repeated six times, resulting in 810 stimulus presentations per participant.

Antisaccade Task We utilized the antisaccade data from the EEGEyeNet dataset [13], which implements the internationally standardized protocol developed by [29]. Each trial began with a central fixation square, which participants were instructed to fixate on for a randomized interval between 1 and 3.5 seconds. Subsequently, a peripheral cue (a dot) appeared either to the left or right of the fixation point. In prosaccade (control) trials, participants were instructed to shift their gaze directly toward the cue. In antisaccade trials, they were instructed to look in the opposite direction, suppressing the reflexive response. The cue remained visible for one second.

Visual Symbol Search This paradigm is also from the EEGEyeNet dataset [13]. The *Visual Symbol Search* (VSS) task is a digital adaptation of the Symbol Search subtest from the WISC-IV and WAIS-III, both standard measures of processing speed [30]. Fifteen stimulus rows are presented simultaneously. In each row, two *target* symbols are followed by five *search* symbols and two response icons labelled "YES" and "NO." For every row, participants decide whether either target appears among the search symbols and register their choice by clicking the appropriate response icon.

Natural Reading To extend the dataset with more naturalistic scenarios, we newly recorded participants during reading tasks. We adapted the protocol from the openly available *ZuCo 2.0* corpus [31], which pairs eye-tracking with EEG during natural reading. Our recordings contain data from participants who read English sentences under two conditions: (i) free, unconstrained reading and (ii) a task-oriented mode requiring detection of a predefined semantic relation in each sentence. We collected these recordings using equipment matching the specifications of the EEGEyeNet dataset to ensure compatibility.

Movie Watching As our second contribution of new recordings, the *Movie Watching* paradigm captures EEG responses in a fully naturalistic viewing context. We recorded participants watching two brief, engaging video excerpts: the bedtime-story scene from *Despicable Me* (MPEG-4, 171 s) and the educational clip *Fun Fractals* (MPEG-4, 163 s). This paradigm represents viewing behaviors not present in the original EEGEyeNet dataset and provides a critical test of scanpath decoding under complex, dynamic visual stimulation. These recordings were collected using the same equipment setup as described in Section 3.3 to maintain consistency across all paradigms.

C.3 Preprocessing and Data Annotation

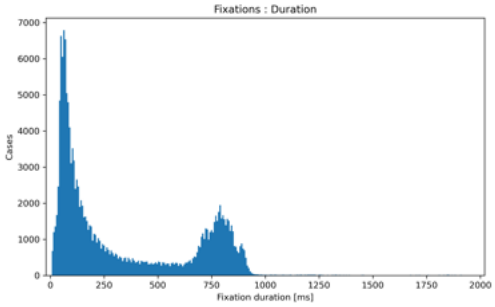
Eye-Tracking Preprocessing Existing literature studying eye movement generally distinguishes between three different ocular events [21]: saccades, fixations, and blinks. *Saccades* are rapid, ballistic eye movements that instantly change the gaze position. Saccade onsets are detected using the eye-tracking software default settings: acceleration larger than $8000°/s^2$, a velocity above 30°/s, and a deflection above 0.1°. *Fixations* are defined as time periods while maintaining the gaze on a single location [18], and *blinks* are considered a special case of fixation, where the pupil diameter is zero.

Electroencephalography Preprocessing EEG data is often contaminated by artifacts produced by environmental factors, e.g., temperature, air humidity, as well as other sources of electromagnetic noise, such as line noise [32] and therefore requires preprocessing in the form of artifact cleaning or artifact correction [33]. Our EEG preprocessing included detecting and interpolating bad electrodes and filtering the data with a 0.1 Hz high-pass filter. With this "minimal" preprocessing pipeline, the preprocessed data still retains ocular activity, which can be used to infer the ocular events (the signal of interest in the present study).

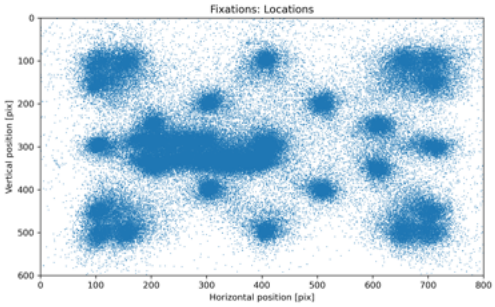
EEG–Eye-Tracking Synchronisation The EEG and eye-tracking streams were temporally aligned with the EYE-EEG toolbox [4], enabling fixation- and saccade-locked EEG analyses. The procedure first detected events common to both recordings (paradigm start and end triggers). A linear regression on these shared markers then refined the eye-tracker time stamps. Synchronisation accuracy was verified by comparing trigger latencies in the two modalities; all deviations were below 2 ms (i.e., one sample).

C.4 Event Characteristics

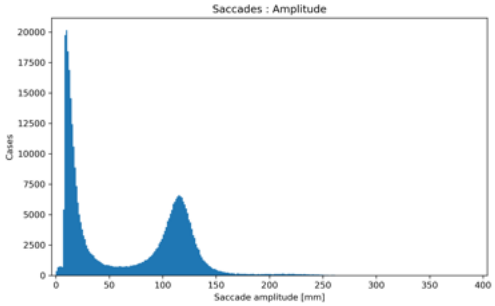
This appendix provides supplementary figures for each dataset. For every experimental paradigm and recording setup, we include (i) distributions of fixation durations, (ii) spatial maps of fixation locations, (iii) histograms of saccade amplitudes, and (iv) angular distributions of saccade directions.



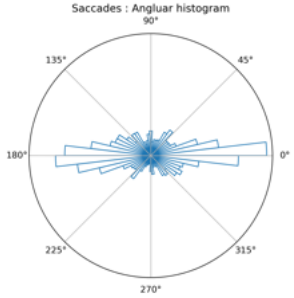
(a) Distribution of the fixation duration



(b) Distribution of the fixation positions

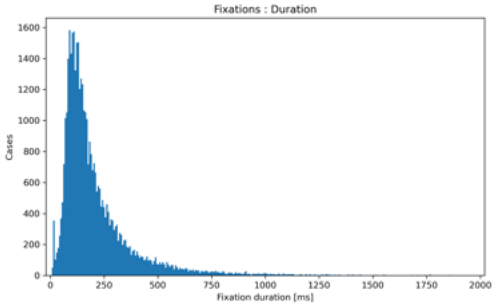


(c) Distribution of the saccade amplitude

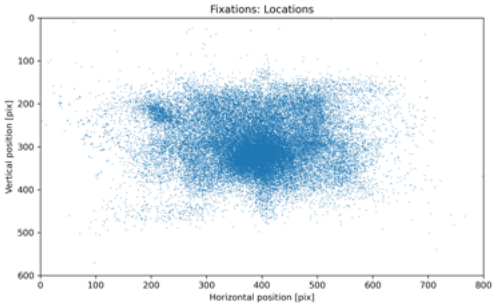


(d) Distribution of the saccade angle

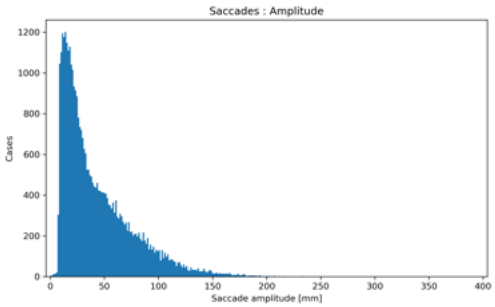
Figure 9: Dots Paradigm



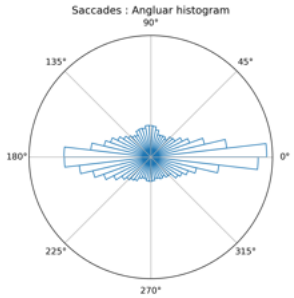
(a) Distribution of the fixation duration



(b) Distribution of the fixation positions

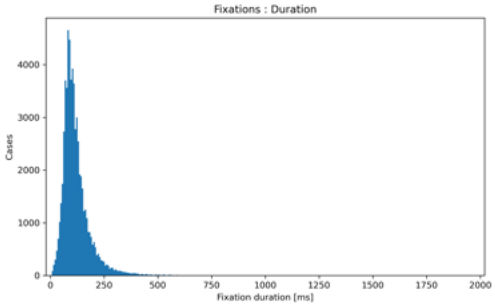


(c) Distribution of the saccade amplitude

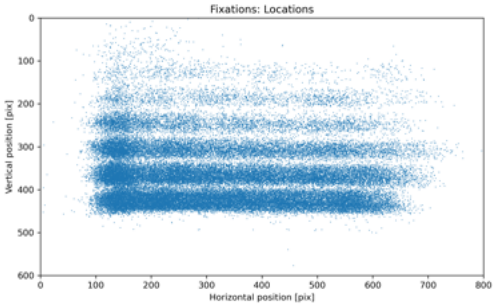


(d) Distribution of the saccade angle

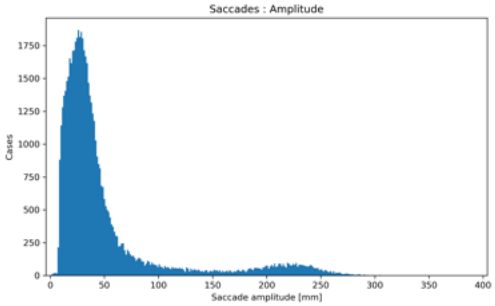
Figure 10: Movie Paradigm



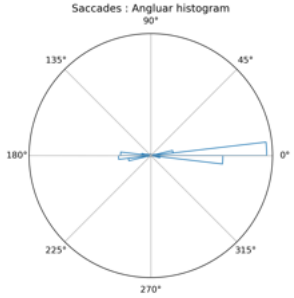
(a) Distribution of the fixation duration



(b) Distribution of the fixation positions

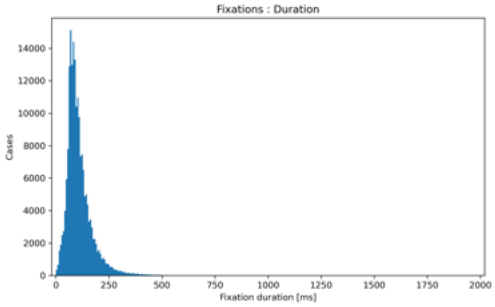


(c) Distribution of the saccade amplitude

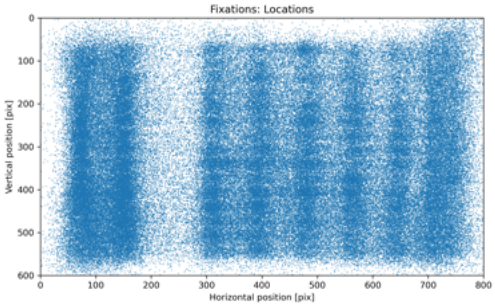


(d) Distribution of the saccade angle

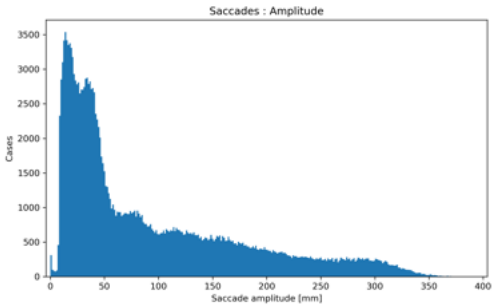
Figure 11: Reading Paradigm



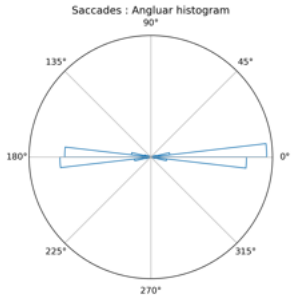
(a) Distribution of the fixation duration



(b) Distribution of the fixation positions

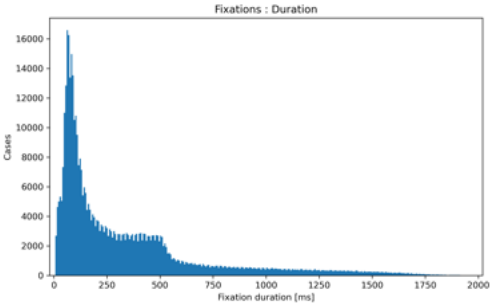


(c) Distribution of the saccade amplitude

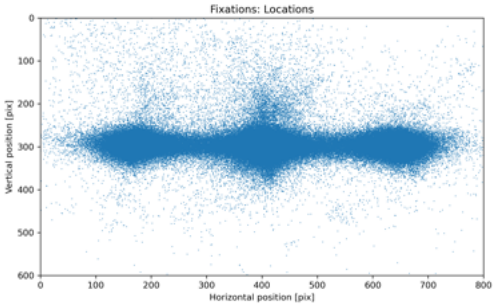


(d) Distribution of the saccade angle

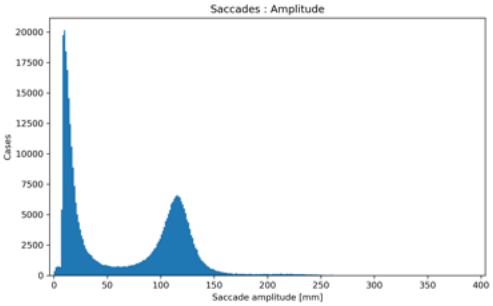
Figure 12: VSS Paradigm



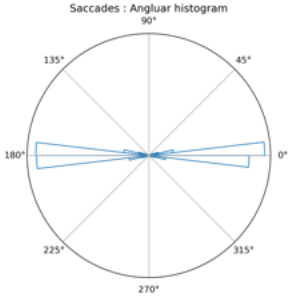
(a) Distribution of the fixation duration



(b) Distribution of the fixation positions



(c) Distribution of the saccade amplitude



(d) Distribution of the saccade angle

Figure 13: ANTI Paradigm

D Generalization Across Hardware

We evaluate generalization between two distinct high-density systems: the EGI Geodesic Hydrocel and the ANT Neuro.

Because the electrode nets for these two systems have different physical layouts and channel numbering, data from one system cannot be directly fed into a model trained on the other. To bridge this gap, we implement a spatial channel mapping based on a nearest neighbor strategy.

Specifically, for each electrode in the source system, we identify the electrode in the target system with the smallest Euclidean distance in 3D scalp space. This establishes a direct mapping where each source channel is assigned to its closest physical counterpart in the target hardware (as illustrated in Figure 14).

While this spatial mapping aligns the physical inputs, the core of our cross-hardware adaptation relies on fine-tuning. We use the spatial map to initialize the model on the target hardware, and then fine-tune it to adapt to the specific characteristics of the new system. Our evaluation follows these steps:

1. **Pretraining:** Train a base model on the source system (e.g., Geodesic Hardware 1).
2. **Spatial Alignment:** Reorder the input channels of the target data (e.g., ANT Neuro Hardware 2) according to the spatial matching to align with the source model's expectations.
3. **Fine-Tuning:** Fine-tune the model on the target system using varying amounts of calibration data. We include a baseline of "0 minutes" of calibration data, which represents pure zero-shot transfer relying entirely on the spatial map.
4. **Evaluation:** Test the fine-tuned model on the target system's held-out test data.

This approach allows us to determine how much target-hardware data is actually needed to achieve strong decoding performance when starting from a spatially aligned, pre-trained model.

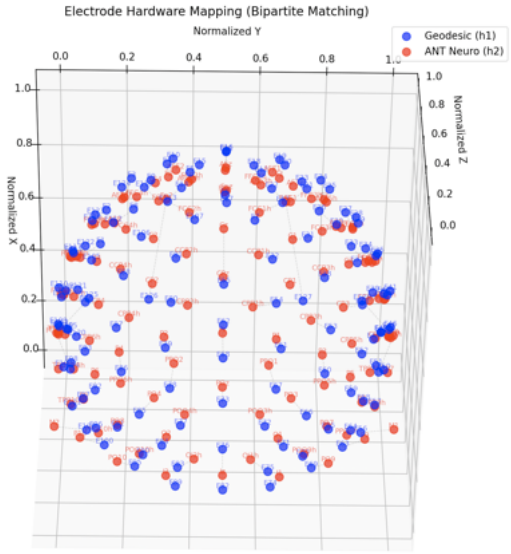


Figure 14: Hardware mapping between the Geodesic system (H1) and ANT Neuro (H2), used to align channels before cross-hardware fine-tuning.

E Mathematical Concepts

For better understanding and clarity, we include an in-depth decomposition of the mathematical concepts behind the loss function design and evaluation metrics used.

E.1 Unified Loss and Matching

Throughout this section we use the same symbols as the main text. Each event has a temporal interval $b = [x_1, x_2]$ (target) or $\hat{b} = [\hat{x}_1, \hat{x}_2]$ (prediction) on the normalised time axis of the segment, an event class $k \in \{\text{fixation}, \text{saccade}, \text{blink}, \text{dummy}\}$, and a screen endpoint $\mathbf{p} = (p^x, p^y) \in \mathbb{R}^2$. For saccades and blinks, $\mathbf{s} = (s^x, s^y)$ is the (true) start position on the screen and $\mathbf{\Delta} = \mathbf{p} - \mathbf{s}$ is the target shift. Predicted counterparts carry a hat: $\hat{\mathbf{\Delta}}$ for the predicted shift and $\hat{\mathbf{p}}$ for the predicted endpoint. Screen coordinates are zero-centred at the orthogonal projection of the eye onto the screen, so $(0, 0)$ is the point straight ahead from the viewer (see Section E.2).

The training objective is a sum of three per-pair terms applied after the matching step,

$$\mathcal{L}_{\text{unified}} \;=\; \mathcal{L}_{\text{CE}} \;+\; \mathcal{L}_{\text{temporal}} \;+\; \mathcal{L}_{\text{spatial}}.$$

$\mathcal{L}_{\text{CE}}$ is averaged over all N predictions (matched slots receive the ground-truth class, unmatched slots the dummy class). $\mathcal{L}_{\text{temporal}}$ and $\mathcal{L}_{\text{spatial}}$ are averaged over the $M \le N$ matched non-dummy pairs only.

E.1.1 Classification loss ($\mathcal{L}_{\text{CE}}$)

A standard categorical cross-entropy supervises the event-type head. For a one-hot target y with $C = 4$ classes and predicted softmax probabilities $\hat{y}$,

$$\text{CE}(y, \hat{y}) \;=\; -\sum_{i=0}^{C-1} y_i \log \hat{y}_i \;=\; -\log \hat{y}_k,$$

where k is the index of the ground-truth class. The dummy class enables a fixed pool of N queries to cover a variable number of events per segment.

E.1.2 Temporal loss ($\mathcal{L}_{\text{temporal}}$)

This term combines an ℓ_1 distance on the boundary coordinates with a generalised-IoU loss on the same intervals,

$$\mathcal{L}_{\text{temporal}}(b, \hat{b}) \;=\; \|b - \hat{b}\|_1 \;+\; \mathcal{L}_{\text{GIoU}}(b, \hat{b}),$$

with the two sub-terms defined below. Both are non-negative and minimised when $\hat{b} = b$.

ℓ_1 boundary loss. The ℓ_1 term penalises the absolute difference in start and end coordinates,

$$\|b - \hat{b}\|_1 \;=\; |x_1 - \hat{x}_1| + |x_2 - \hat{x}_2|.$$

This drives the predicted boundaries numerically close to the target.

Generalised-IoU loss. Standard temporal IoU is

$$\text{IoU}(b, \hat{b}) \;=\; \frac{\text{Inter}(b, \hat{b})}{\text{Union}(b, \hat{b})},$$

where $\text{Inter}(b, \hat{b}) = \max\big(0, \min(x_2, \hat{x}_2) - \max(x_1, \hat{x}_1)\big)$ and $\text{Union}(b, \hat{b}) = (x_2 - x_1) + (\hat{x}_2 - \hat{x}_1) - \text{Inter}(b, \hat{b})$. IoU vanishes for non-overlapping intervals, providing no gradient. Generalised IoU fixes this by adding the smallest enclosing interval $b_{\text{enc}} = [\min(x_1, \hat{x}_1), \max(x_2, \hat{x}_2)]$ with length $|b_{\text{enc}}| = \max(x_2, \hat{x}_2) - \min(x_1, \hat{x}_1)$,

$$\text{GIoU}(b, \hat{b}) \;=\; \text{IoU}(b, \hat{b}) \;-\; \frac{|b_{\text{enc}}| - \text{Union}(b, \hat{b})}{|b_{\text{enc}}|} \;\in\; [-1, 1].$$

GIoU is a similarity (1 at perfect overlap, -1 for far-away intervals), so we use it as a loss in the standard form

$$\mathcal{L}_{\text{GIoU}}(b, \hat{b}) \;=\; 1 - \text{GIoU}(b, \hat{b}) \;\in\; [0, 2].$$

The penalty $(|b_{\text{enc}}| - \text{Union})/|b_{\text{enc}}|$ measures how much of the enclosing interval is not covered by the union of b and $\hat{b}$, providing an informative gradient even when the two intervals do not overlap.

E.1.3 Spatial loss ($\mathcal{L}_{\text{spatial}}$)

The spatial loss supervises the predicted endpoint of every matched non-dummy event. We instantiate it along two axes already introduced in the main text and exercised in the ablation: the *coordinate system* (Cartesian or polar) and the *loss form* (Euclidean or disaggregated ℓ_1). All quantities are expressed on the screen plane in millimetres unless stated otherwise, and we treat the screen as zero-centred at the orthogonal projection of the eye.

From parametrisation to predicted endpoint. Each query emits a 2-d output $\hat{\mathbf{u}} = (\hat{u}_1, \hat{u}_2)$ from its spatial head. We map this output to a Cartesian shift $\hat{\boldsymbol{\Delta}} = (\hat{\Delta}^x, \hat{\Delta}^y)$ via

$$\hat{\boldsymbol{\Delta}} \;=\; \begin{cases} (\hat{u}_1,\, \hat{u}_2) & \text{Cartesian,} \\ \hat{r}\,(\cos\hat{\theta},\, \sin\hat{\theta}), \quad (\hat{u}_1, \hat{u}_2) = (\hat{r}, \hat{\theta}) & \text{Polar.} \end{cases}$$

The predicted endpoint $\hat{\mathbf{p}}$ on the screen is then anchored class-conditionally,

$$\hat{\mathbf{p}} \;=\; \begin{cases} \hat{\boldsymbol{\Delta}} & \text{(fixation; anchor is the screen centre } \mathbf{0}\text{),} \\ \mathbf{s} + \hat{\boldsymbol{\Delta}} & \text{(saccade or blink; anchor is the true start } \mathbf{s}\text{).} \end{cases}$$

At training time the saccade/blink anchor is always the ground-truth start, matching the implementation; the same convention is used at evaluation time for the per-step error metrics defined in Section E.2.

Euclidean form. The unified spatial objective scores the predicted endpoint against the target endpoint with a screen-normalised Euclidean distance,

$$\mathcal{L}_{\text{Euclidean}}(\mathbf{p}, \hat{\mathbf{p}}) \;=\; \sqrt{\left(\frac{\hat{p}^{\,x} - p^x}{W}\right)^2 + \left(\frac{\hat{p}^{\,y} - p^y}{H}\right)^2}. \tag{2}$$

Per-axis normalisation by the screen width W and height H makes the loss dimensionless and display-independent. It couples the two spatial axes through a single endpoint distance, so a unit of error along x is comparable to a unit of error along y in fractions of the corresponding screen extent.

Disaggregated ℓ_1 form. The disaggregated variant scores the two output components independently in the parametrisation space of $\hat{\mathbf{u}}$, foregoing the rotation invariance of the endpoint distance in exchange for separable axis-wise terms. With $\mathbf{u} = (u_1, u_2)$ the corresponding ground-truth coordinates,

$$\mathcal{L}^{\ell_1}_{\text{spatial}}(\mathbf{u}, \hat{\mathbf{u}}) \;=\; \frac{|\hat{u}_1 - u_1|}{N_1} \;+\; \frac{d_2(\hat{u}_2, u_2)}{N_2},$$

with normalisers and per-coordinate distance:

- Cartesian: $(N_1, N_2) = (W, H)$, $d_2(a, b) = |a - b|$.
- Polar: $(N_1, N_2) = (D, 1)$, $d_2(\hat{\theta}, \theta) = \big|\text{atan2}\big(\sin(\hat{\theta} - \theta), \cos(\hat{\theta} - \theta)\big)\big| \;\in\; [0, \pi]$, the circular angular distance.

The circular distance is equivalent to $\arccos\big(\cos(\hat{\theta} - \theta)\big)$ and avoids the spurious 2π jump that a plain $|\hat{\theta} - \theta|$ would create at the wrap-around.

Total spatial objective. Let $\mathcal{M} = \{(i_k, \sigma(i_k))\}_{k=1}^{M}$ be the set of M matched pairs returned by the Hungarian assignment for which the matched ground-truth class is fixation, saccade, or blink (dummy queries do not contribute to $\mathcal{L}_{\text{spatial}}$). With class weights $w_{\ell_k} \in \{w_{\text{fix}}, w_{\text{sac}}, w_{\text{blink}}\}$ depending on the matched class, the spatial loss aggregated over a batch is

$$\mathcal{L}_{\text{spatial}}^{\text{batch}} \;=\; \frac{1}{M} \sum_{k=1}^{M} w_{\ell_k} \; \mathcal{L}_{\text{spatial}}\big(\mathbf{p}_{i_k}, \hat{\mathbf{p}}_{\sigma(i_k)}\big),$$

where $\mathcal{L}_{\text{spatial}} \in \{\mathcal{L}_{\text{Euclidean}}, \mathcal{L}_{\text{spatial}}^{\ell_1}\}$. Dummy queries contribute to $\mathcal{L}_{\text{CE}}$ only.

E.1.4 Matching

Since the decoder emits a fixed set of N event predictions while a segment contains $M \leq N$ ground-truth events, we need a one-to-one alignment before computing the per-pair losses. Following DETR [26], we solve a bipartite matching problem with the Hungarian algorithm: we find an injective map $\sigma\colon \{1, \ldots, M\} \to \{1, \ldots, N\}$ minimising the total assignment cost,

$$\hat{\sigma} \;=\; \arg\min_{\sigma} \sum_{i=1}^{M} \mathcal{C}_{\text{match}}\big(y_i, \hat{y}_{\sigma(i)}\big).$$

The remaining $N - M$ unmatched queries are assigned to the dummy class. As an alternative, we also implement a non-crossing dynamic-programming assignment that additionally enforces temporal monotonicity, $i < i' \Rightarrow \sigma(i) < \sigma(i')$ (used when the corresponding flag is enabled).

Per-pair matching cost. The matching cost is a weighted sum of the same per-pair quantities used in $\mathcal{L}_{\text{unified}}$, grouped into a single temporal term $\mathcal{C}_{\text{time}}$ that mirrors $\mathcal{L}_{\text{temporal}}$ in the main text,

$$\mathcal{C}_{\text{match}} \;=\; \lambda_{\text{cls}}\, \mathcal{C}_{\text{cls}} \;+\; \mathcal{C}_{\text{time}} \;+\; \lambda_{\text{feat}}\, \mathcal{C}_{\text{spatial}}, \qquad \mathcal{C}_{\text{time}} \;=\; \lambda_{\text{bbox}}\, \mathcal{C}_{\text{L1}} \;+\; \lambda_{\text{giou}}\, \mathcal{C}_{\text{GIoU}},$$

with components

- $\mathcal{C}_{\text{cls}} \;=\; -\hat{p}_i(k)$, the negated softmax probability of the target class k.
- $\mathcal{C}_{\text{L1}} \;=\; \|b - \hat{b}\|_1$, identical to the temporal ℓ_1 term.
- $\mathcal{C}_{\text{GIoU}} \;=\; -\text{GIoU}(b, \hat{b})$. We drop the additive constant (the matcher uses $-\text{GIoU}$ rather than $1 - \text{GIoU}$, since constants do not affect the assignment).
- $\mathcal{C}_{\text{spatial}}$ follows the same formula as $\mathcal{L}_{\text{spatial}}$ above, evaluated on the candidate prediction's class-conditional spatial head selected by the ground-truth class of the target.

Using the same per-pair quantities for matching and for the training loss guarantees that every candidate pair is scored consistently across the two stages.

E.2 Evaluation Metrics

In our evaluation, we compare predicted scanpaths against ground truth scanpaths sampled at 500 time steps over 1 second. The goal is to quantify how well the predicted scanpath aligns with the true eye movement data using meaningful metrics. We use the following metrics:

1. **F1 Score**
   Purpose: Classification of eye events. Captures discrete event labeling quality
2. **Visual Distance for Fixation Errors**
   Purpose: Spatial error of fixation points. Accounts for perceptual visual angle at fixation points
3. **Visual Distance for Saccades/Blinks Errors**
   Purpose: Spatial error of saccade and blink endpoints. Accounts for perceptual visual angle at saccade/blink endpoints
4. **Dynamic Time Warping (DTW)**
   Purpose: Sequence-level alignment of scanpaths. This metrics handles temporal misalignment and speed differences in scanpath sequences

Below we describe the metrics used in detail, including the visual distance formulation and the Dynamic Time Warping (DTW) metric.

E.2.1 F1 Score

To assess classification performance of eye events (fixation, saccade, blink), we use the F1 score. For binary classification, F1 is defined as the harmonic mean of precision and recall:

$$F_1 = 2 \cdot \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

For our case of multi-class classification, we compute the F1 score independently for each class by treating it as the positive class and all others as negative. This yields one F1 score per class, which can be averaged to produce an overall metric.

E.2.2 Per-Step Metrics

Upon successful segmentation, we can compute event-specific metrics. We measure all errors using unified visual distance.

Visual Distance. To quantify spatial discrepancies between predicted and true gaze points, we use a visual distance metric that measures the angular difference between two points as perceived from the eye, rather than their raw Euclidean distance on the screen plane.

Coordinate convention. We assume that screen coordinates (x, y) are zero-centred at the orthogonal projection of the eye onto the screen. In other words, the origin $(0, 0)$ on the screen is the point that lies straight ahead from the viewer, and f denotes the perpendicular eye-to-screen distance in millimetres. With this convention, $\sqrt{x^2 + y^2 + f^2}$ is exactly the 3-D distance from the eye to the screen point (x, y), which is what the law-of-cosines derivation below requires.

Let

$$\mathbf{x} = (x_1, y_1), \quad \mathbf{y} = (x_2, y_2)$$

be two zero-centred points on the screen plane, and let $f =$ DISTANCE_TO_EYE (in mm) be the perpendicular eye-to-screen distance. We compute the 3-D distances from the eye to each point,

$$d_1 = \sqrt{x_1^2 + y_1^2 + f^2}, \quad d_2 = \sqrt{x_2^2 + y_2^2 + f^2},$$

and the Euclidean distance between the points on the screen plane,

$$D = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

Using the law of cosines, the visual angle $\Delta\theta$ subtended at the eye by the two points satisfies

$$\cos \Delta\theta = \frac{d_1^2 + d_2^2 - D^2}{2 d_1 d_2},$$

and we clamp the right-hand side to $[-1, 1]$ for numerical stability before taking the arccosine. The visual distance metric is then

$$\text{VisualDistance}(\mathbf{x}, \mathbf{y}) = \Delta\theta = \arccos\left(\frac{d_1^2 + d_2^2 - D^2}{2 d_1 d_2}\right).$$

This metric better reflects human perceptual differences in gaze location than raw pixel distances. As mentioned above, when computing per-step errors we distinguish between fixations and saccades/blinks.

Fixation Error. For a fixation event with target endpoint $\mathbf{p}$ (the true gaze point) and predicted endpoint $\hat{\mathbf{p}}$ obtained from the spatial head as described in Appendix E.1.3, the per-event error is the visual distance between the two,

$$\text{FixationError} \;=\; \text{VisualDistance}(\hat{\mathbf{p}}, \mathbf{p}).$$

Saccade and Blink Error. Saccades and blinks are characterised by a true start $\mathbf{s}$ and a true endpoint $\mathbf{p}$ on the screen, with target shift $\boldsymbol{\Delta} = \mathbf{p} - \mathbf{s}$. The model emits a predicted shift $\hat{\boldsymbol{\Delta}}$ which we anchor at the *true* start position to obtain the predicted endpoint,

$$\hat{\mathbf{p}} \;=\; \mathbf{s} + \hat{\boldsymbol{\Delta}}.$$

This anchoring convention is the same as that used during training (Appendix E.1.3), so what we report at evaluation time is the residual error of the predicted shift with respect to the ground-truth start, not a compound error that also reflects mistakes in detecting the start position. The per-event error is then the visual distance between the predicted and true endpoints,

$$\text{SaccadeError} \;=\; \text{VisualDistance}(\hat{\mathbf{p}}, \mathbf{p}),$$

and analogously for blinks. This unified metric captures the key spatial aspect of saccades and blinks as deviations from the true starting point on the screen.

E.2.3 Dynamic Time Warping (DTW)

While per-step metrics quantify pointwise spatial errors, scanpaths are sequences that can vary in timing and length. Using simple raw point-to-point distance over time does not account for temporal distortions such as:

- Small shifts or delays in timing of fixations/saccades
- Variable saccade duration or blink intervals
- Local accelerations or decelerations in scanpath speed

DTW addresses these by finding an optimal alignment ("warping") between predicted and true scanpath sequences, minimizing overall visual distance cost under monotonicity and continuity constraints.

Formally, for sequences

$$X = (x_1, \ldots, x_N), \quad Y = (y_1, \ldots, y_M)$$

with points $x_n, y_m \in \mathbb{R}^2$, we build the cost matrix

$$C(n, m) = \text{VisualDistance}(x_n, y_m)$$

DTW finds a warping path $p = (p_1, \ldots, p_L)$, $p_l = (n_l, m_l)$ minimizing the total cost. This alignment enables matching corresponding gaze events despite temporal shifts, producing a more perceptually meaningful similarity score than raw frame-wise distance.

F Error Analysis

In the main paper (Figure 5) we report the three spatial-error views aggregated across paradigms. The same three views are reproduced below as a 2×3 grid per paradigm, using the model trained jointly on all five paradigms (Hardware 1). Each panel reuses the metric and styling of the corresponding main-paper figure, so panels are directly comparable across paradigms; "ALL" repeats the pooled view from the main paper for reference. We report errors in degrees of visual angle (eye-to-screen distance 680 mm).

Spatial Distribution of Fixation Errors. Figure 15 shows, for each paradigm, the 2D density of fixation prediction error $(\Delta x, \Delta y) = \widehat{g} - g$, where $\widehat{g}$ and g are the decoded and ground-truth gaze positions of correctly classified fixation events. Densities are normalised per panel so the grayscale levels reflect the *shape* of the error distribution rather than absolute event counts (which differ by an order of magnitude across paradigms). Three patterns are visible. (i) Highly structured paradigms (ANTI, Dots) produce the tightest central concentration, since fixation targets are constrained to a small set of screen locations. The discrete satellite peaks visible for Dots reflect the regular grid of stimulus positions: when the model's prediction snaps to a neighbouring dot, a systematic offset of constant magnitude is introduced. (ii) Naturalistic and free-viewing paradigms (Movie, Reading, WISC) produce broader, more diffuse error distributions. (iii) The dispersion is anisotropic: WISC and to a lesser degree Reading exhibit clear vertical elongation, indicating larger errors along the vertical axis than along the horizontal – consistent with the well-documented asymmetry between horizontal and vertical EEG-based gaze decoding.

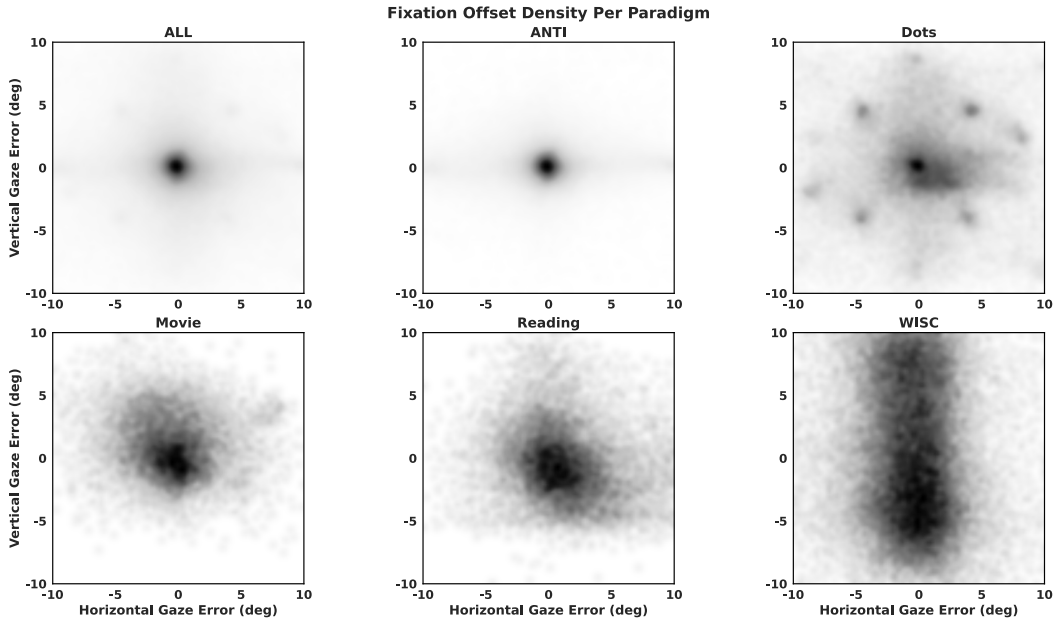


Figure 15: Fixation offset density per paradigm. Each panel shows the 2D density of decoded-vs-ground-truth fixation offset (deg of visual angle), per-panel min–max normalised. ANTI and Dots produce tight central concentrations (the regular satellite peaks for Dots reflect the grid of stimulus positions); Reading and especially WISC show vertical elongation, consistent with weaker vertical decoding than horizontal.

Directional Bias. Figure 16 reproduces the angular view of the main paper per paradigm: the radial coordinate is the mean Cartesian endpoint error (in degrees of visual angle) for saccade and blink events whose true direction falls into a given θ bin. The horizontal-pinch / vertical-bulge shape of the pooled curve is present in every paradigm, confirming that this directional asymmetry is a property of EEG-based gaze decoding and not an artefact of any single dataset. ANTI and Reading have the smallest overall radial extent; the ANTI shape is also the most circular, consistent with its strong horizontal task structure equalising the per-direction sampling. Dots produces a markedly spiky

profile because its stimuli sit on a sparse grid of cardinal and oblique directions, causing the per-bin means to alternate between well- and under-sampled directions. Movie and WISC have the largest overall extent; WISC additionally shows the strongest vertical excess (roughly 4° at $\theta = \pm\pi/2$ vs. 2–3° at $\theta = 0, \pm\pi$), again pointing to vertical asymmetry as the dominant source of residual error in unstructured viewing.



Figure 16: Mean endpoint error (deg of visual angle) per true saccade direction, polar plot, per paradigm. The horizontal-pinch / vertical-bulge pattern of the pooled (ALL) curve is present in every paradigm. WISC shows the strongest vertical excess; Dots produces a spiky profile because its stimuli sample only a sparse set of cardinal and oblique directions.

Impact of Saccade Amplitude. Figure 17 extends the main-paper amplitude view per paradigm. In each panel, the line shows mean endpoint error as a function of true saccade amplitude (binned in 1° steps), and the grey histogram on the secondary axis shows the count of events per bin. Both axes are auto-scaled *within each panel* so the within-paradigm shape of the curve and its amplitude distribution are easy to read. Across all paradigms the error grows approximately monotonically with amplitude, with the steepness of the growth and the location of the count peak being paradigm-dependent: Reading's amplitudes concentrate below 5° (consistent with within-line text saccades), Dots is approximately uniform up to 20° (driven by the fixed stimulus grid), and Movie has a long-tailed, smoothly-decaying amplitude distribution typical of free viewing. Movie also exhibits the largest absolute errors at long amplitudes, while ANTI, Reading and WISC stay below $\sim 8^\circ$ throughout.

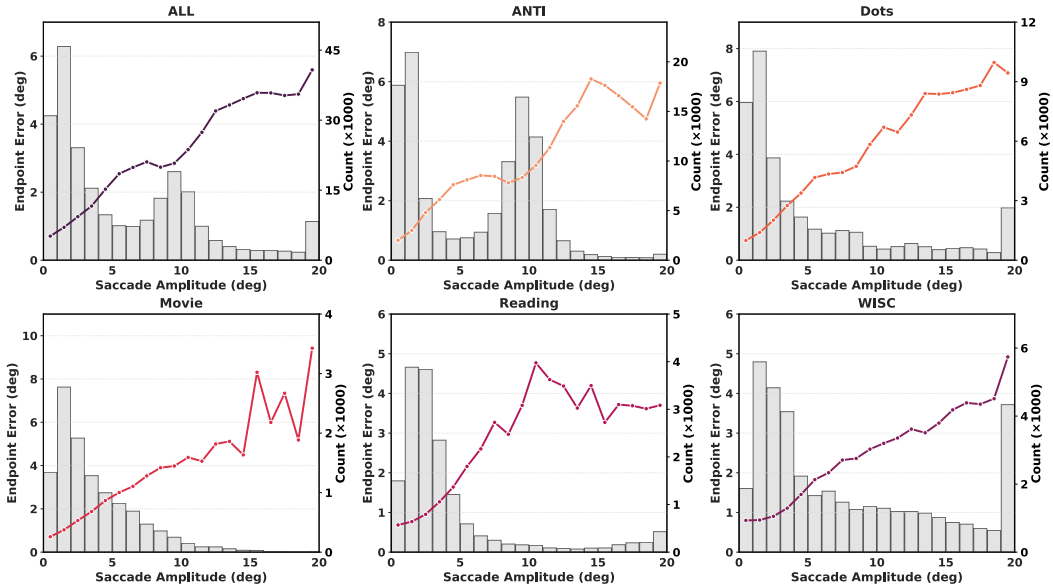


Figure 17: Endpoint error (deg) as a function of true saccade amplitude, per paradigm. Mean-error and count axes are auto-scaled within each panel so the within-paradigm shape (approximately monotonic in every panel) and the paradigm-specific amplitude distribution are visible at a glance.

G Additional Decoding Examples Across Paradigms

We provide here additional examples for continuous decoding trajectories for each paradigm included in the benchmark.

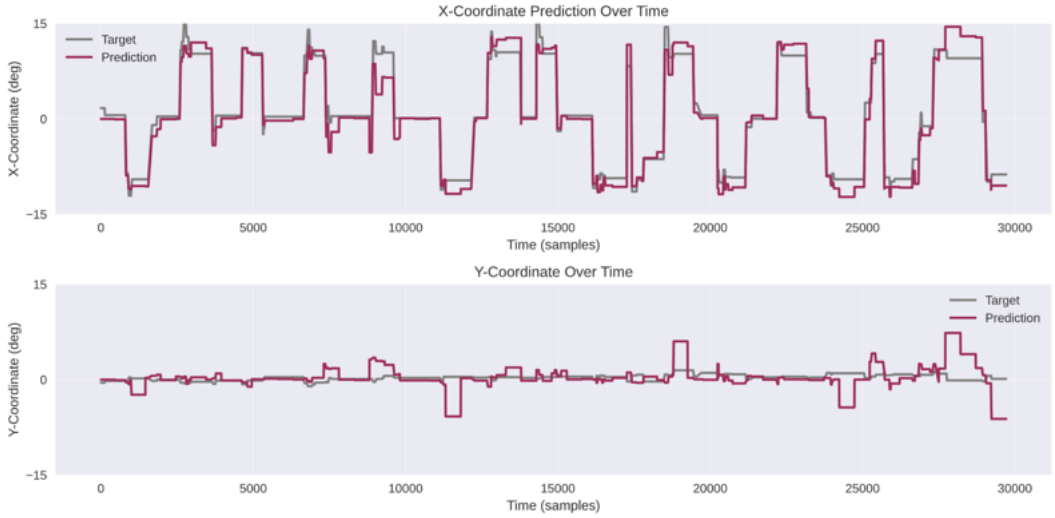


Figure 18: Continuous decoding trajectory for the **Anti-saccade** paradigm. The model accurately captures the sharp, discrete saccadic jumps and stable fixations across the horizontal (X) axis.

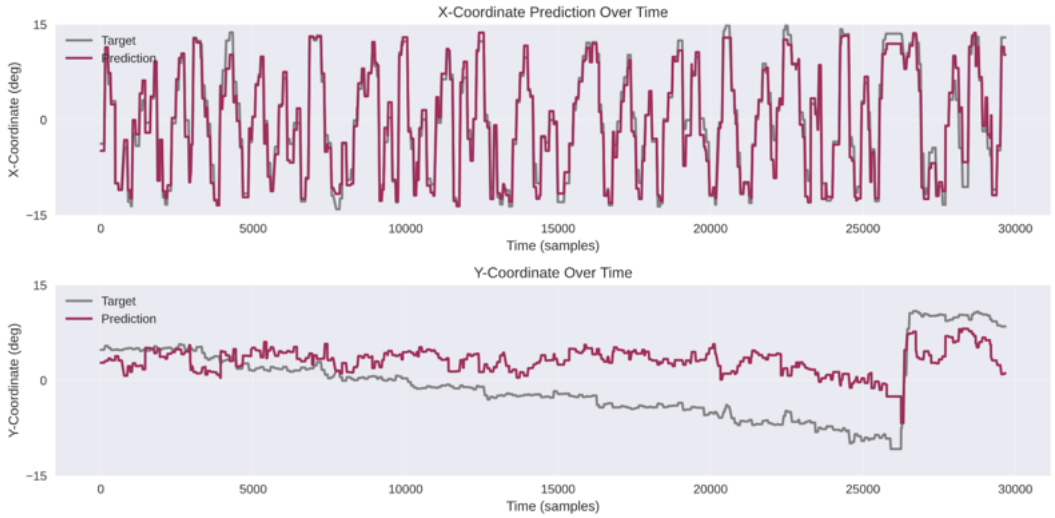


Figure 19: Continuous decoding trajectory for the **Visual Search Task (VSS)**. This paradigm is characterized by dense, high-frequency ocular movements in both dimensions as the subject scans the search array.

The visualizations demonstrate that the unified decoder successfully learns the distinct types of different eye movement tasks. In the **Reading** paradigm (Figure 20), the model reproduces the characteristic "sawtooth" pattern on the horizontal axis, representing progressive fixations across a line followed by a rapid return saccade. In the **Anti-saccade** task (Figure 18), the trajectories show high-amplitude, discrete jumps that align closely with the target, indicating high precision in structured environments. Conversely, the naturalistic and search-based tasks highlight the current limitations of EEG-based gaze decoding. In the **Visual Search (VSS)** task (Figure 19), while the model captures the high-frequency search dynamics, there is a visible drift in the Y-coordinate

reconstruction. This reflects our earlier finding that absolute vertical position is significantly harder to decode from brain activity than horizontal movement. In the **Movie** paradigm (Figure 21), the movements are more fluid and less structured, showing that while the model follows the general gaze trend, it occasionally struggles with the fine-grained nuances of unconstrained viewing.

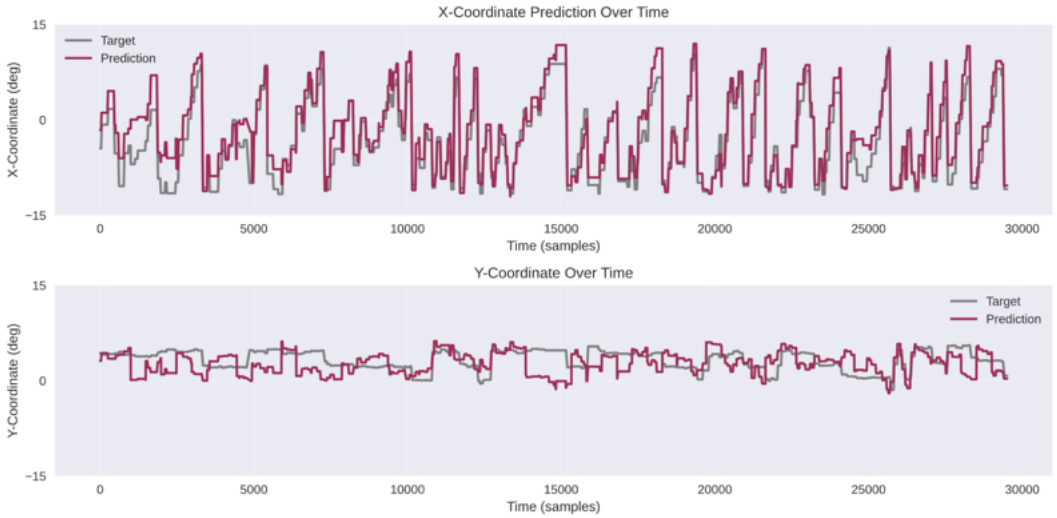


Figure 20: Continuous decoding trajectory for the **Reading** task. Note the distinctive "sawtooth" pattern in the X-coordinate, representing progressive fixations across a line of text followed by a rapid return saccade to the start of the next line.



Figure 21: Continuous decoding trajectory for the **Movie** watching paradigm, showing more naturalistic transitions compared to the task-based paradigms.

Finally, in Figure 22, we also show decoding examples in screen coordinates for both the unified model and the baseline.

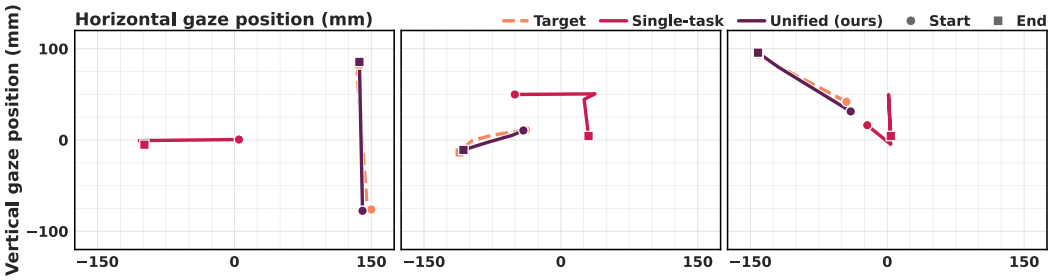


Figure 22: Examples of scanpath reconstructions from our new one-shot Unified Decoder in cases where Single-Task pipelines aren't effective.

H Calibration (Fine-Tuning) Process

To evaluate how much subject-specific data is needed for optimal performance, we follow a structured fine-tuning pipeline. As shown in Figure 23, the process consists of three main stages:

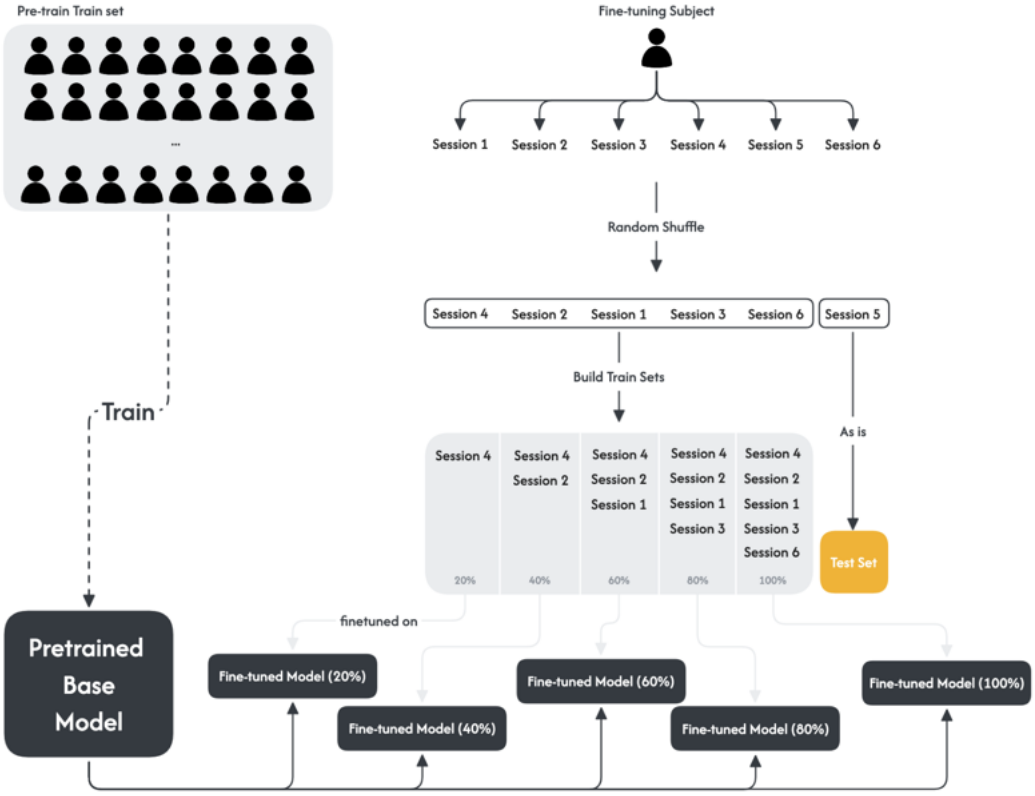


Figure 23: The calibration pipeline. A pretrained base model is fine-tuned on increasing amounts of session-based data from a single subject to evaluate how quickly accuracy recovers after a domain or hardware shift.

Pre-training: A base model is first trained on a large-scale dataset containing multiple subjects and paradigms. This allows the model to learn general EEG-to-gaze mappings.

Session Shuffling and Splitting: For a new "fine-tuning subject," we collect multiple recording sessions. To ensure the evaluation is robust, we randomly shuffle these sessions. One session is held out entirely as a **Test Set** ("As is"), while the remaining sessions are used to build incremental training sets.

Incremental Fine-Tuning: We build five training subsets in 20% increments (e.g., using one session for the 20% model, up to five sessions for the 100% model, in each using at most 5 minutes of data). This incremental approach allows us to measure the "learning curve" of the calibration process and identify the point where adding more subject-specific data yields diminishing returns.