EYELIGNR: Aligning Gaze and Text via Contrastive Learning for Improved Reading Scanpath Prediction

Ekta Sood*
University of Colorado Boulder
ekta.sood@colorado.edu

Simon Tannert*
University of Stuttgart
simon.tannert@ims.uni-stuttgart.de

Prajit Dhar
University of Marburg
dhar@uni-marburg.de

Meghdut Sengupta
University of Hannover
m.sengupta@ai.uni-hannover.de

Enrica Troiano
HK3Lab
enrica.troiano@hk3lab.ai

Sidney K. D'Mello
University of Colorado Boulder
sidney.dmello@colorado.edu

Abstract

Modeling the interaction between gaze dynamics and linguistic structure remains a central challenge in scanpath prediction. Existing approaches either incorporate linguistic information in ways that do not fully respect the temporal progression of reading, or ignore linguistic structure altogether. Their effectiveness is further limited by the scarcity and heterogeneity of available eye-tracking data. To address these issues, we introduce EYELIGNR, a framework for spatiotemporal scanpath prediction that integrates linguistic information at the representation level. EYELIGNR conditions language representations on fixation order and aligns gaze and text within a shared embedding space through contrastive learning. Experiments on five reading corpora show that EYELIGNR generalizes effectively across reading conditions, improving over prior state of the art on multiple scanpath metrics (e.g., up to NLD ∼57%, NLL ∼27%, and FLA ∼59%).

1 Introduction

Scanpaths are sequences of eye movements over visual stimuli, and when recorded on text, they provide a signal of reading behavior. What makes these data particularly valuable is that they capture three aspects that come together in human text processing (Kintsch, 1992, 1994): the linguistic features of the text being read (e.g., lexical and semantic properties), the spatial trajectories of the eyes (e.g., forward and backward movements), and the temporal unfolding of those movements (e.g., order and duration of word fixations). This combination makes gaze information relevant for several areas of natural language processing (NLP). It directly motivates the task of *scanpath prediction*, which aims to estimate the sequence of word fixations during reading, while also supporting downstream applications that leverage gaze to improve NLP models (Mishra et al., 2017; Hollenstein and Zhang, 2019; Sood et al., 2020; Khurana et al., 2023, i.a.). To date, however, jointly modeling the linguistic, spatial, and temporal aspects of reading remains difficult.

Researchers have explored several methods to equip scanpath models with linguistic information (Deng et al., 2024; Bolliger et al., 2025). These methods typically align gaze encodings with pretrained language representations by mapping the sequence of eye fixations (e.g., "Sally to the store went") onto a canonical left-to-right textual order ("Sally went to the store"). In doing so, they do not directly encode the actual progression of reading, which often involves nonlinear phenomena like word skipping or backtracking. The temporal dynamics of eye movements have not been fully exploited either: most studies consider only word fixation order while overlooking duration (Deng et al., 2023; Bolliger et al., 2023), which is a key signal of human attentional allocation and language understanding. Lastly, data scarcity poses additional challenges (Stebakov and Pershin, 2025). Training robust predictors requires large amounts of eye-tracking data, but such data are collected under varying experimental protocols, making standardization difficult and limiting the models' generalization abilities.

Recent work has begun to address some of these problems with ScanEZ, an architecture that explicitly incorporates fixation duration into gaze representations, and that mitigates data scarcity through self-supervised modeling (Sood et al., 2025). Notably, ScanEZ achieves state-of-the-art results on

*These authors contributed equally.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 42147–42160
October 24-29, 2026 ©2026 Association for Computational Linguistics

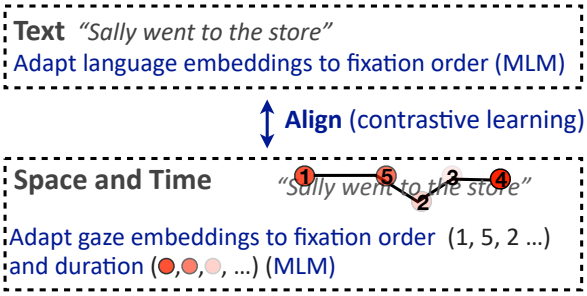


Figure 1: Modeling the linguistic, spatial, and temporal dimensions of reading under data scarcity. In EYELIGNR, fixation-ordered language and gaze representations, including fixation duration, are obtained via modality-specific masked modeling objectives and aligned contrastively for scanpath prediction.

scanpath prediction without using any linguistic input, raising the question of whether language information can yield further improvements. We address this question here by expanding on the modeling approach of ScanEZ with a principled integration of gaze and text.

This paper presents EYELIGNR[1], a contrastive gaze–text alignment framework for spatiotemporal scanpath prediction. EYELIGNR constructs textual representations that preserve the temporal sequence of reading by adapting a pretrained language encoder with masked language modeling (MLM) on fixation-ordered token sequences. The architecture combines this adapted language encoder with a Transformer-based encoder for eye movements, learning a shared multimodal embedding space through contrastive training. This design models spatial trajectories, temporal dynamics (including fixation duration), and linguistic context within a unified framework (cf. Figure 1). The self-supervised pretraining and contrastive multimodal alignment further enable EYELIGNR to effectively learn gaze–text representations and remain robust to limited and heterogeneous eye-tracking data.

In sum, we advance scanpath prediction with:

(1) a language representation strategy based on MLM that fine-tunes a pretrained encoder on fixation-ordered token sequences;

(2) a contrastive gaze–text alignment approach that injects linguistic information into spatiotemporal scanpath representations.

Based on these contributions, we test the hypothesis that gaze-informed text embeddings improve scanpath prediction. We show that this is indeed the case: our approach surpasses models that use no language representations (scanpath metrics improve by as much as 58.6%) or assume a canonical left-to-right word order (82.6%). The gains hold on five datasets spanning controlled and naturalistic reading settings, providing evidence that EYELIGNR generalizes across reading conditions.

[1]Code will be provided upon request.

2 Related Work

A growing body of research in NLP uses eye movements during reading, either as prediction targets (Li and Rudzicz, 2021; Srivastava, 2022) or as auxiliary supervisory signals for downstream tasks (Barrett and Hollenstein, 2020; **?**; Bondar et al., 2025). Still, leveraging gaze representations poses two related challenges.

First, data scarcity. Limited eye-tracking data constrain the scalability of scanpath models, especially in the absence of standardized evaluation protocols. Shubi et al. (2025) address this gap by introducing EyeBench, a benchmark for predictive modeling from gaze that harmonizes datasets and standardizes evaluation across different generalization regimes; Stebakov and Pershin (2025) tackle a complementary problem with a unified evaluation framework that benchmarks existing models across datasets, reading conditions, and psycholinguistic phenomena. Beyond evaluation-focused efforts, the use of synthetic scanpaths has emerged as a promising response to the data scarcity bottleneck. In particular, ScanEZ (Sood et al., 2025) proposes pretraining on scanpaths generated from the cognitively grounded reading model E-Z Reader (Reichle et al., 2003), before fine-tuning on smaller human eye-tracking datasets. By framing scanpath prediction as a self-supervised sequence modeling problem, ScanEZ learns continuous spatiotemporal fixation trajectories via a masked reconstruction objective, without reliance on large labeled datasets.

Second, language–gaze alignment. Researchers typically address the mismatch between word order and the order of reading fixations at the architectural level. For instance, Eyettention encodes the stimulus text in the left-to-right word order and the scanpath following the reading order, aligning the two streams via local cross-attention to predict the next fixation (Deng et al., 2023). ScanDL and ScanDL 2.0 (Bolliger et al., 2023, 2025) formulate scanpath generation as a diffusion process condi-

tioned on the stimulus sentence, using frozen pretrained BERT embeddings as textual conditioning while learning gaze dynamics in a joint embedding space. In these models, language representations remain fixed and are aligned to gaze through attention mechanisms or embedding reindexing, limiting their ability to model reading-order-specific language-gaze interactions.

In this work, we retain ScanEZ's Transformer-based encoder for spatiotemporal gaze modeling and its masked fixation reconstruction objective for self-supervised learning. We augment ScanEZ with language through contrastive learning, an approach that provides a data-efficient alternative to sequence-level supervision and that has been shown effective for constructing multimodal gaze representations (Jindal and Manduchi, 2023; Zhao et al., 2024). Crucially, we shift the focus from the architectural level to the representation level. Rather than aligning frozen textual embeddings through attention mechanisms or token reindexing, we directly obtain a reading-aware text embedding space by fine-tuning the language encoder parameters on fixation-ordered token sequences.

3 Task Definition and the EYELIGNR Architecture

A scanpath consists of a sequence of fixations over a piece of text, which may be a page or a stimulus of a different textual granularity depending on the dataset (Deng et al., 2023). Each fixation is represented by a triplet (x, y, t), where x and y denote the continuous spatial coordinates of the fixation location and t is the fixation duration (the temporal order of reading is given, implicitly, by the order of fixations in the sequence). We cast scanpath prediction as a masked sequence modeling task: given a partially masked scanpath, the model predicts the (x, y, t) values of the masked fixations.

To address this task, we build EYELIGNR using three main components: an encoder to process eye-movement sequences, an encoder to process text, and a component for contrastive learning.

Eye-Movement Encoder. We represent gaze sequences with the Transformer-based encoder of ScanEZ. This encoder processes input sequences of shape (B, T, D), where B is the batch size, T is the sequence length (up to 200 fixations), and D is the feature dimensionality (i.e., 3 dimensions corresponding to the spatiotemporal coordinates x, y, t). To handle variable-length sequences, we use attention masking that prevents the model from attending to padded positions.

Language Encoder. For text encoding, we use a pretrained RoBERTa-base model (Liu et al., 2019). The model outputs contextualized representations for each token in an input sequence of shape (B, S, H), where S is the sequence length (up to 512 tokens) and H is the hidden dimension (768).

Contrastive Component (Sequence Pooling and Projection Heads). Similar to Radford et al. (2021), we aggregate variable-length sequences into fixed-size representations using masked mean pooling. For the ScanEZ encoder output:

$$\mathbf{z}_{\text{gaze}} = \frac{\sum_{t=1}^{T} m_t \mathbf{h}_t}{\sum_{t=1}^{T} m_t} \tag{1}$$

where $\mathbf{h}_t$ is the hidden state at time step t and $m_t \in \{0, 1\}$ is the corresponding mask value. Similarly, for the language encoder:

$$\mathbf{z}_{\text{text}} = \frac{\sum_{s=1}^{S} a_s \mathbf{h}_s}{\sum_{s=1}^{S} a_s} \tag{2}$$

where $a_s \in \{0, 1\}$ indicates whether token s is part of the actual text (not padding). We then use linear layers that map pooled representations into a shared 512-dimensional embedding space:

$$\mathbf{e}_{\text{gaze}} = \text{normalize}(W_{\text{gaze}} \mathbf{z}_{\text{gaze}}) \tag{3}$$

$$\mathbf{e}_{\text{text}} = \text{normalize}(W_{\text{text}} \mathbf{z}_{\text{text}}). \tag{4}$$

where "normalize" denotes L2 normalization onto the unit hypersphere. The projection heads are discarded after contrastive training.

4 Training Pipeline

This section describes our approach to learning reading-aware language and gaze representations. We start by introducing fixation-order encoding and then present our training framework.

4.1 Fixation-Order Encoding

For each eye-movement sequence, we construct a fixation-ordered text representation by reordering the tokens according to the temporal sequence of fixations in the scanpath. Given a text with tokens $w_1, w_2, \ldots, w_n$ in left-to-right order and a scanpath with fixations at positions $p_1, p_2, \ldots, p_m$, we create the fixation-ordered sequence: $T_{\text{fixation}} =$

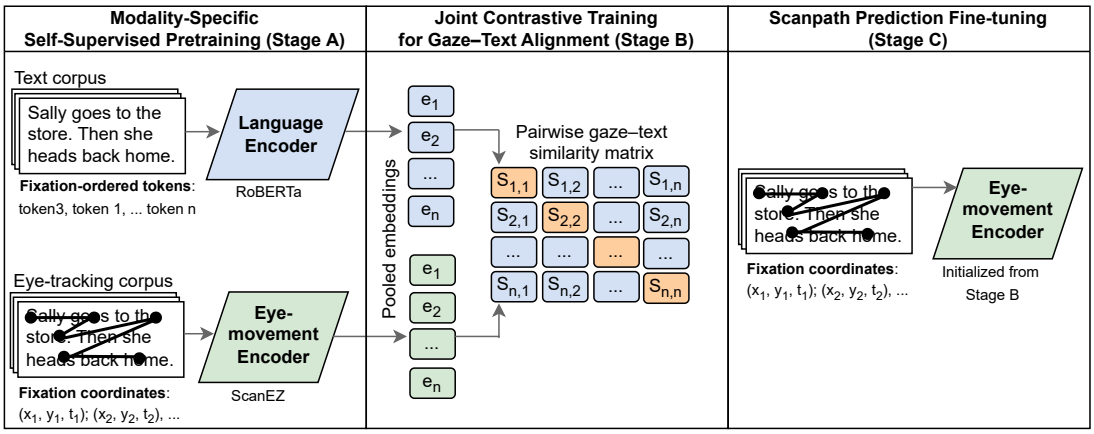


Figure 2: Training pipeline of EYELIGNR. Eye-movement and language encoders are separately pretrained with self-supervised objectives (Stage A). Their pooled representations are then aligned via contrastive learning in a shared embedding space, using a pairwise similarity matrix computed between pooled gaze and text embeddings (Stage B). Lastly, the contrastive projection heads are discarded, and the gaze encoder and regression head are fine-tuned for scanpath prediction (Stage C).

$[w_{p_1}, w_{p_2}, \ldots, w_{p_m}]$, where each w_{p_i} is the token fixated at time step i. This representation captures reading phenomena such as refixation (tokens may appear at multiple time steps) and skipping (tokens may not appear at all in the gaze sequence).

4.2 Training Procedure

Our training pipeline includes three stages illustrated in Figure 2: (A) modality-specific pretraining, (B) contrastive pretraining for multimodal alignment, and (C) scanpath prediction fine-tuning.

Stage A: Modality-Specific Pretraining. We start by pretraining the eye-movement encoder and the language encoder to learn modality-specific representations of reading behavior.

For the eye-movement prediction task, we initialize the ScanEZ encoder from the checkpoint in Sood et al. (2025) and continue training it using a masked modeling approach adapted for sequential eye-movement data. Following BERT's masking strategy (Devlin et al., 2019), we randomly mask a subset of the input fixation triplets and train the model to reconstruct them: for each sequence, we randomly select a subset $\mathcal{M}$ of fixation positions and mask their (x, y, t) values. The ScanEZ encoder processes the partially masked sequence, and the regression head predicts the masked triplets. The training objective is:

$$\mathcal{L}_{\text{pretrain}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} (X_i - \hat{X}_i)^2 \qquad (5)$$

where X_i are the true fixation coordinates and duration at position i, and $\hat{X}_i$ are the model's predictions. With pretraining, the encoder is encouraged to learn bidirectional contextual representations of reading behavior, as it must use surrounding fixations to infer masked positions.

Regarding the language model, we fine-tune RoBERTa on *fixation-ordered* sequences with a masked language modeling objective. Fine-tuning exposes the model to the statistical patterns of fixation-ordered text and encourages representations that are robust to token reordering. We initialize the model from the pretrained RoBERTa-base checkpoint and train using RoBERTa's standard objective (Liu et al., 2019): we randomly mask 15% of tokens in the fixation-ordered sequence and learn to predict the masked tokens based on their fixation-order context.

Stage B: Joint Contrastive Pretraining. In the second stage, we jointly optimize the pretrained ScanEZ and RoBERTa encoders to align eye-movement representations with their corresponding textual contexts. The goal is to promote alignment at the representation level, without enforcing explicit token-to-fixation correspondences or architectural coupling.

The training uses a symmetric contrastive loss based on the InfoNCE objective (Oord et al., 2018). For a batch of N paired eye-movement–text examples, we first obtain pooled representations and project them into the shared embedding space. We then compute the temperature-scaled cosine simi-

larity matrix:

$$S_{ij} = \frac{\mathbf{e}^{(i)}_{\text{gaze}} \cdot \mathbf{e}^{(j)}_{\text{text}}}{\tau} \tag{6}$$

where τ is a temperature parameter that controls the sharpness of the distribution. We set $\tau = 0.05$ based on preliminary experiments showing this value provides stable training and good separation between positive and negative pairs.

The contrastive loss is computed in both directions (gaze-to-text and text-to-gaze) and averaged:

$$\mathcal{L}_{\text{e2t}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(S_{ii})}{\sum_{j=1}^{N} \exp(S_{ij})} \tag{7}$$

$$\mathcal{L}_{\text{t2e}} = -\frac{1}{N}\sum_{j=1}^{N} \log \frac{\exp(S_{jj})}{\sum_{i=1}^{N} \exp(S_{ij})} \tag{8}$$

$$\mathcal{L}_{\text{contrastive}} = \frac{\mathcal{L}_{\text{e2t}} + \mathcal{L}_{\text{t2e}}}{2} \tag{9}$$

This formulation treats the corresponding eye-movement and text representations of each example as a positive pair, while representations from different examples serve as negative pairs. With batch size N, the model learns to distinguish the correct pairing from $N - 1$ negative pairs.

Stage C: Scanpath Fine-tuning. In the final stage, we adapt the contrastively learned representations to scanpath prediction. After removing the contrastive projection heads, the ScanEZ encoder (with the regression head) is fine-tuned on the task, using the same masked prediction objective as in Equation 5. This stage allows the model to exploit the gaze–text alignment learned during contrastive training while optimizing for scanpath prediction.

See Appendix A.1 for more training details.

5 Experimental Setup

We evaluate EYELIGNR using 5-fold cross validation. For each fold, we split the data into 80%/20% for training/testing. Following past work (e.g., Deng et al., 2023; Sood et al., 2025), we prevent data leakage by ensuring that train and test splits are fully disjoint with respect to both textual stimuli and participants, i.e., all test examples involve previously unseen texts and unseen readers. This evaluation protocol is applied across multiple eye-tracking corpora using standard scanpath prediction metrics, as described below. Additional train/test split strategies and metrics are discussed in Appendix A.2 and A.3.

5.1 Eye-tracking Datasets

To test model robustness under data distributional shifts, we use five corpora that vary in stimulus granularity (e.g., sentences, full articles), genre, task demands, participant populations, scanpath lengths and coordinate distributions (due to, e.g., screen geometries and data collection protocols).

CELER L1, ZuCo 1.0, and EML (henceforth CELER, ZuCo, and EML) are the benchmarks used in the evaluation of ScanEZ. CELER (Berzak et al., 2022) is a large-scale corpus suitable for fine-grained word-level fixation modeling, comprising data from 365 participants who read English sentences from newswire and Wikipedia. Like Sood et al. (2025), we focus on data from the 69 English L1 speakers. ZuCo (Hollenstein et al., 2018) contains eye-tracking recordings from 12 participants reading isolated sentences and Wikipedia articles. Following Sood et al. (2025), we focus on sentences from the natural reading task. EML (Caruso et al., 2022) contains scanpaths typical of real-world learning, obtained from 147 fluent English readers engaged with long-form educational texts (five ∼1,000-word passages across multiple pages).

SB-SAT (Ahn et al., 2020) includes long-text scanpaths, obtained from 95 undergraduates in a standardized multi-page reading comprehension test (SAT practice contexts and questions).

ROAMM (Sun et al., 2025) is a multimodal (eye-tracking and EEG) corpus of naturalistic reading and attention lapses. With data from 44 participants reading five full-length Wikipedia articles, this corpus provides beyond-sentence data, with realistic variation in topic and reader engagement.

Statistics of the data extracted from the corpora are summarized in Table 1. For preprocessing, we first derive raw scanpaths as timestamped fixation sequences, with the (x, y, t) coordinates aligned with the corresponding text tokens; we then harmonize scanpaths across participants and hardware configurations by normalizing fixation duration (dividing t by the maximum duration within each scanpath) and spatial coordinates (dividing x and y by the screen resolution used during recording).[2]

[2]Normalization leaves some degree of cross-dataset heterogeneity due to differences in screen layout (e.g., y varies across corpora depending on line breaks and paging).

Dataset	Part.	Sent.	Text Source	# Scan.	X (px)	Y (px)
CELER	69	156	News / Wikipedia	8514	666.84 ± 446.64	539.06 ± 167.50
ZuCo	12	700	Wikipedia / Sentiment	2887	329.34 ± 145.22	139.32 ± 66.11
EML	147	239	Educational texts	5808	901.25 ± 276.26	394.43 ± 192.41
SB-SAT	95	118	Standardized test (SAT)	240	477.80 ± 257.93	329.06 ± 175.20
ROAMM	44	514	Wikipedia	1720	921.75 ± 400.58	459.52 ± 307.05

Table 1: Statistics of the eye-tracking data used in our experiments. Part./Sent.: participants/sentences (for ROAMM and SB-SAT, we identify sentence-final punctuation in tokenized texts). # Scan: total number of scanpaths. X/Y: scanpath coordinates in screen-centered pixel space (mean ± standard deviation); negative values are due to dataset-specific coordinate conventions.

5.2 Evaluation Metrics

We consider five traditional scanpath metrics that capture separate prediction dimensions:

- *NLL_x, NLD_x, NLL_t*, i.e., the negative log-likelihood of predicted horizontal fixation coordinates, Normalized Levenshtein Distance of predicted horizontal fixation coordinates, and the negative log-likelihood of predicted fixation durations.

- *Fixation Duration Accuracy (FDA)* measures the model's accuracy in predicting how long readers fixate each word, calculated as:

$$d = |\hat{T} - T| \tag{10}$$

$$FDA = 1 - \frac{d}{\max(d)} \tag{11}$$

  where $\hat{T}$ and T are the predicted and ground-truth fixation durations, respectively.

- *Fixation Location Accuracy (FLA)* evaluates spatial prediction accuracy, i.e., whether the model correctly predicts which words are fixated, calculated as:

$$d = (\hat{X} - X)^2 + (\hat{Y} - Y)^2 \tag{12}$$

$$FLA = 1 - \frac{\sqrt{d}}{\max(\sqrt{d})} \tag{13}$$

  where $\hat{X}$ and $\hat{Y}$ are the predicted coordinates and X and Y are the ground-truth coordinates.

Better performance corresponds to lower values of NLL_x, NLD_x, and NLL_t, and higher values of FDA and FLA.

We further use Root Mean Squared Error (RMSE) to compare performance between models on the x, y, and t dimensions. Paired significance tests are performed on per-scanpath differences, using Bonferroni-adjusted p-values to account for multiple comparisons, and effect sizes are reported using Cohen's d for paired samples ($|d| < 0.2$: negligible magnitude, $0.2 \leq |d| < 0.5$: small, $0.5 \leq |d| < 0.8$: moderate, $|d| \geq 0.8$: large).

Lastly, we investigate the quality of predictions using MultiMatch (Dewhurst et al., 2012), which measures similarity between predicted and ground-truth scanpaths in terms of shape, direction, length, position, and duration. Higher MultiMatch scores along these dimensions indicate higher similarity between predictions and ground truth at the sequence level.

6 Results

To test the hypothesis that gaze-informed text representations improve scanpath prediction, we assess EYELIGNR against ScanEZ (Section 6.1) and a "Token-Ordered" baseline (Section 6.2). ScanEZ provides a language-ablated version of EYELIGNR; comparisons with this model therefore isolate the contribution of linguistic information over purely spatiotemporal modeling.[3] Comparing with Token-Ordered additionally tests the effect of adapting language to the unfolding of gaze: this model is trained using the same architecture and data as EYELIGNR, but without adjusting the pretrained language embeddings to reading order.

6.1 Language Improves Scanpath Prediction

EYELIGNR surpasses ScanEZ on the spatial metrics NLL_x, NLD_x, and fixation-location accuracy (Table 2). Gains are visible across datasets, particularly on CELER and EML: on CELER, NLL_x is reduced by 26.6% (1.524→1.118), NLD_x by 56.8% (.421→.182), and FLA increases by

[3] ScanEZ is the current state-of-the-art scanpath predictor on the same datasets used in our evaluation. Therefore, benchmarking against ScanEZ provides a comparison between our model and earlier architectures.

Data	Model	$NLL_x\downarrow$	$NLD_x\downarrow$	$NLL_t\downarrow$	FDA↑	FLA↑
CELER	ScanEZ	1.524 ± 0.04	0.421 ± 0.02	**1.240 ± 0.02**	0.650 ± 0.00	0.413 ± 0.01
	Token-Ordered	2.503 ± 0.28	0.446 ± 0.01	2.028 ± 0.02	0.613 ± 0.01	0.391 ± 0.03
	EYELIGNR	**1.118 ± 0.09**	**0.182 ± 0.01**	1.599 ± 0.01	**0.693 ± 0.00**	**0.655 ± 0.01**
ZuCo	ScanEZ	1.140 ± 0.07	0.365 ± 0.00	1.334 ± 0.12	0.718 ± 0.03	0.623 ± 0.02
	Token-Ordered	1.573 ± 0.07	0.420 ± 0.01	1.803 ± 0.09	0.643 ± 0.02	0.292 ± 0.03
	EYELIGNR	**1.053 ± 0.09**	**0.335 ± 0.01**	**1.218 ± 0.10**	**0.756 ± 0.03**	**0.662 ± 0.04**
EML	ScanEZ	0.996 ± 0.02	0.616 ± 0.01	1.083 ± 0.03	0.804 ± 0.01	0.534 ± 0.01
	Token-Ordered	1.253 ± 0.02	0.416 ± 0.00	1.686 ± 0.03	0.650 ± 0.01	0.530 ± 0.00
	EYELIGNR	**0.747 ± 0.09**	**0.330 ± 0.00**	**1.039 ± 0.02**	**0.818 ± 0.00**	**0.727 ± 0.00**
SB-SAT	ScanEZ	2.904 ± 0.176	0.406 ± 0.00	2.342 ± 0.11	0.820 ± 0.01	0.738 ± 0.01
	Token-Ordered	3.434 ± 0.16	0.461 ± 0.00	3.375 ± 0.049	0.493 ± 0.01	0.460 ± 0.01
	EYELIGNR	**2.704 ± 0.15**	**0.346 ± 0.01**	**2.288 ± 0.12**	**0.825 ± 0.00**	**0.805 ± 0.02**
ROAMM	ScanEZ	2.713 ± 0.11	0.392 ± 0.01	1.394 ± 0.07	0.839 ± 0.01	0.736 ± 0.01
	Token-Ordered	3.324 ± 0.12	0.462 ± 0.00	2.368 ± 0.10	0.554 ± 0.01	0.443 ± 0.00
	EYELIGNR	**2.524 ± 0.161**	**0.341 ± 0.01**	**1.321 ± 0.06**	**0.847 ± 0.01**	**0.809 ± 0.01**

Table 2: Scanpath prediction results (mean ± standard deviation). CELER, ZuCo and EML correspond to the datasets used for the ScanEZ evaluation, as reported in Sood et al. (2025).

58.6% (.413→.655); on EML, NLL_x drops by 25% (.996→.747) and NLD_x by 46.4% (.616→.330), with FLA rising by .193 points (.534→.727).

EYELIGNR generally outperforms ScanEZ on fixation-duration metrics as well. The only exception is CELER, where ScanEZ attains lower NLL_t (1.240) despite having lower FDA (0.650) than EYELIGNR. This outcome indicates that ScanEZ achieves better distributional fit on CELER, but EYELIGNR produces more accurate pointwise predictions. This behavior is especially plausible given CELER's constrained temporal structure (i.e., single-line reading and low variance in fixation durations), which favors models with well-calibrated uncertainty over sharper point estimates.

EYELIGNR therefore yields superior spatial trajectory modeling and, in most cases, improved temporal modeling relative to ScanEZ. The exception found on CELER is unsurprising, because contrastive language–gaze alignment was not designed to optimize fixation-duration prediction. Nonetheless, the improvements on all other datasets indicate that linguistic grounding can provide a beneficial inductive bias for temporal modeling under naturalistic reading conditions. Taken together, these results establish EYELIGNR as a new state-of-the-art approach for scanpath prediction.

Sources of Spatial Performance Gains. As a diagnostic analysis, we investigate the contribution of horizontal (x) and vertical (y) predictions to the gains in spatial accuracy. To this end, we examine whether EYELIGNR has an advantage over ScanEZ in terms of RMSE.[4]

An asymmetry emerges between the x and y dimensions, suggesting that FLA reflects accurate horizontal predictions more than vertical ones. On x, EYELIGNR performs significantly better than ScanEZ across all datasets (e.g., $p < .001$, d=-2.917 on SB-SAT). Conversely, on y, ScanEZ generally performs better (e.g., on ZuCo, d(RMSE)=1.021). We take this finding as evidence that our training procedure successfully pushed EYELIGNR to focus on language-related phenomena: EYELIGNR captures within-line reading dynamics, which are linked to linguistic processing, better than vertical across-line transitions, which are mainly determined by text layout (like screen constraints) rather than text content.

6.2 Accounting for Reading Order Matters

EYELIGNR outperforms the Token-Ordered baseline (Table 2). Improvements are observed across all datasets and metrics. For instance, spatial performance is boosted by 82.6% on ROAMM (FLA: 0.443→0.809); on SB-SAT, the temporal evaluation has a 32.2% reduction in NLL_t and a 67.3% increase in fixation-duration accuracy. Interestingly, Token-Ordered underperforms both EYELIGNR and ScanEZ. This suggests that the use of *token-ordered* language embeddings can be detrimental not only relative to models based on *fixation-ordered* embeddings, but even relative to models that incorporate no linguistic representations at all.

[4]Full dataset-by-dimension RMSEs are in Appendix A.3.

Data	Model	Shape	Dir	Len	Pos	Dur
CELER	Token-Ordered	0.986 ± 0.02	0.652 ± 0.01	0.984 ± 0.01	0.827 ± 0.03	0.579 ± 0.01
	EYELIGNR	**0.993 ± 0.00**	**0.977 ± 0.00**	**0.988 ± 0.00**	**0.971 ± 0.00**	**0.787 ± 0.01**
ZuCo	Token-Ordered	0.976 ± 0.00	0.778 ± 0.01	0.971 ± 0.00	0.665 ± 0.01	0.645 ± 0.01
	EYELIGNR	**0.980 ± 0.00**	**0.917 ± 0.00**	**0.972 ± 0.00**	**0.929 ± 0.00**	**0.783 ± 0.01**
EML	Token-Ordered	0.973 ± 0.00	0.704 ± 0.00	0.973 ± 0.00	0.818 ± 0.03	0.642 ± 0.01
	EYELIGNR	**0.983 ± 0.00**	**0.908 ± 0.02**	**0.977 ± 0.00**	**0.938 ± 0.01**	**0.789 ± 0.04**
SB-SAT	Token-Ordered	0.971 ± 0.00	0.740 ± 0.01	0.969 ± 0.00	0.742 ± 0.01	0.380 ± 0.02
	EYELIGNR	**0.979 ± 0.00**	**0.915 ± 0.00**	**0.971 ± 0.00**	**0.922 ± 0.00**	**0.741 ± 0.01**
ROAMM	Token-Ordered	0.972 ± 0.00	0.740 ± 0.00	0.969 ± 0.00	0.734 ± 0.02	0.499 ± 0.02
	EYELIGNR	**0.979 ± 0.00**	**0.912 ± 0.00**	**0.972 ± 0.00**	**0.925 ± 0.00**	**0.779 ± 0.00**

Table 3: MultiMatch similarity results (mean and standard deviation). For all metrics, the higher the better. Dir: direction. Len: length. Pos: position. Dur: Duration.

Emergent Temporal Gains from Fixation-Order Alignment. Comparisons based on RMSE corroborate the advantage of EYELIGNR over the Token-Ordered model. Along the x dimension, advantages are more pronounced than those observed earlier relative to ScanEZ (e.g., on SB-SAT: d=-6.378). More importantly, EYELIGNR achieves significantly lower RMSE than Token-Ordered on the duration dimension t with consistently large effects (e.g., SB-SAT: d=-2.761). In other words, by learning to predict better trajectories, EYELIGNR also approximates more realistic temporal patterns. This finding is consistent with prior work, which posits that fixation duration co-varies with spatial reading dynamics such as regressions and word skipping (Pollatsek et al., 2008).

Sequence-Level Evaluation of Fixation-Order Alignment. The metrics considered so far point to the benefits of fine-tuning on fixation-ordered language sequences and contrastive alignment. Yet, they evaluate accuracy locally, at the level of individual fixation coordinates or durations, without assessing the global quality of predictions. We thus use MultiMatch to assess entire scanpaths across several dimensions of sequence similarity.

EYELIGNR outperforms the Token-Ordered baseline across the five dimensions of shape, direction, length, position, and duration (Table 3). Particularly large gains are observed in directional similarity. For example, on CELER, scores increase from 0.652 with Token-Ordered to 0.977 with EYELIGNR, and on SB-SAT from 0.740 to 0.915. Similarly, for positional similarity, on CELER scores rise from 0.827 to 0.971, and on SB-SAT from 0.742 to 0.922.

Since directional similarity reflects patterns of saccades and regressions, the large gains of EYELIGNR support our hypothesis: aligning language representations with fixation order enables us to better model eye movements through text. In contrast, language embeddings derived from the canonical word order (as in Token-Ordered) introduce a mismatch between textual and gaze representations that disrupts the coherence of predicted scanpath sequences. Improvements in positional similarity further indicate that predicted fixations land closer to the locations observed in human scanpaths, suggesting that EYELIGNR more accurately captures which words are fixated, skipped, or revisited. This finding corroborates that EYELIGNR improves horizontal trajectory modeling, which largely reflects within-line reading dynamics.

Finally, gains in duration similarity (e.g., Δ(Dur)=0.361 on SB-SAT) suggest that fixation-order alignment produces more realistic temporal patterns, consistent with psycholinguistic evidence showing that fixation duration co-varies with spatial reading behavior, including regressions and refixations triggered by processing difficulty (Pollatsek et al., 2008).

7 Conclusion

We introduce EYELIGNR, a framework that injects linguistic information into a spatiotemporal scanpath model through contrastive gaze–text alignment. Building on masked fixation modeling, we adapt the language encoder on fixation-ordered token sequences that match the non-monotonic order of natural reading, and align the resulting text embeddings with scanpath representations through contrastive learning. In doing so, we address a methodological question left open by prior

work, that is, how to leverage linguistic information for scanpath prediction in a way that complements high-performing eye-movement representations learned from gaze data, while accounting for the spatial and temporal aspects of reading.

Compared to well-established modeling approaches (i.e., both language-agnostic and token-ordered baselines), EYELIGNR achieves improvements on multiple scanpath metrics (e.g., NLD, NLL, and FLA; CELER: $\sim$57%, $\sim$27%, $\sim$59%; EML: $\sim$46%, $\sim$25%, $\sim$36%), setting a strong benchmark across reading conditions, including long-form educational and learning settings. Experimental results highlighted an asymmetry between horizontal and vertical predictions, likely due to cross-corpus variation in page layout. In the future, we will extend the current approach to better model vertical scanpath patterns.

8 Limitations

EYELIGNR is evaluated on multiple English reading corpora, but not across languages, scripts, or interfaces (e.g., mobile, VR), where fixation behavior and layout statistics may differ substantially. Moreover, our approach assumes access to reliable token-to-fixation alignment and consistent preprocessing across corpora; in practice, alignment quality varies with eye-tracking experimental protocols and annotation pipelines, and residual alignment noise can affect both fixation-order language adaptation and contrastive alignment.

Regarding our experimental results, we point out that while EYELIGNR improves core scanpath metrics, gains are not uniform across all objectives. In particular, the trade-off on fixation-duration likelihood in one experimental setting (i.e., NLL_t on CELER) indicates that better spatial and sequence-level alignment does not automatically translate into better calibrated duration likelihoods. The improvements are also not evenly distributed across scanpath dimensions: y is comparatively harder to model under distributional shift, as it exhibits large variance across corpora (e.g., single-line sentences versus multi-page passages), statistics of vertical eye transitions are less consistent across datasets.

Our findings suggest potential relevance for future psycholinguistic research. In particular, the benefits of fixation-order language adaptation indicate that the model may capture aspects of within-line reading dynamics linked to linguistic processing, such as regressions and word skipping. These phenomena are central to theories of human reading and eye-movement control. Yet, the present work does not directly test cognitive or linguistic hypotheses, nor does it establish direct relationships between learned representations and human text processing mechanisms. Future work could therefore investigate whether fixation-aware multimodal representations such as those learned by EYELIGNR can support analyses of reading behavior that are more closely informed by cognitive and psycholinguistic theories.

9 Ethical Considerations

This work did not go through an ethical committee. However, we ensured compliance with ethical principles in data usage and methodology. For the pretraining of the gaze embedding, we used synthetic eye trajectories (no real individuals can be reconstructed from them). For the scanpath prediction task, we relied on existing datasets described in peer-reviewed publications, in which individuals are thoroughly anonymized.

Acknowledgements

This research was supported by the National Science Foundation (DRL 1920510 and 2019805). The opinions expressed are those of the authors and do not represent views of the funding agency.

References

Seoyoung Ahn, Conor Kelton, Aruna Balasubramanian, and Greg Zelinsky. 2020. Towards Predicting Reading Comprehension From Gaze Behavior. In *ACM Symposium on Eye Tracking Research and Applications*, ETRA '20 Short Papers, pages 1–5, New York, NY, USA. Association for Computing Machinery.

Maria Barrett and Nora Hollenstein. 2020. Sequence labelling and sequence classification with gaze: Novel uses of eye-tracking data for Natural Language Processing. *Language and Linguistics Compass*, 14(11):1–16.

Yevgeni Berzak, Chie Nakamura, Amelia Smith, Emily Weng, Boris Katz, Suzanne Flynn, and Roger Levy. 2022. CELER: A 365-Participant Corpus of Eye Movements in L1 and L2 English Reading. *Open Mind*, 6:41–50.

Lena Bolliger, David Reich, Patrick Haller, Deborah Jakobi, Paul Prasse, and Lena Jäger. 2023. ScanDL: A diffusion model for generating synthetic scanpaths on texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15513–15538, Singapore. Association for Computational Linguistics.

Lena Bolliger, David Reich, and Lena Ann Jäger. 2025. Scandl 2.0: A generative model of eye movements in reading synthesizing scanpaths and fixation durations. *Proceedings of the ACM on Human-Computer Interaction*, 9(3):1–29.

Anna Bondar, David Robert Reich, and Lena Ann Jäger. 2025. AlEYEgnment: Leveraging Eye-Tracking-While-Reading to align language models with human preferences. In *Proceedings of the First International Workshop on Gaze Data and Natural Language Processing*, pages 58–70, Varna, Bulgaria. INCOMA Ltd., Shoumen, BULGARIA.

Megan Caruso, Candace E. Peacock, Rosy Southwell, Guojing Zhou, and Sidney K. D'Mello. 2022. Going deep and far: Gaze-based models predict multiple depths of comprehension during and one week following reading. *International Educational Data Mining Society*.

Shuwen Deng, Paul Prasse, David Reich, Tobias Scheffer, and Lena Jäger. 2024. Fine-tuning pre-trained language models with gaze supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–224, Bangkok, Thailand. Association for Computational Linguistics.

Shuwen Deng, David Reich, Paul Prasse, Patrick Haller, Tobias Scheffer, and Lena Ann Jäger. 2023. Eyettention: An attention-based dual-sequence model for predicting human scanpaths during reading. *Proceedings of the ACM on Human-Computer Interaction*, 7(ETRA):1–24.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Richard Dewhurst, Marcus Nyström, Halszka Jarodzka, Tom Foulsham, Roger Johansson, and Kenneth Holmqvist. 2012. It depends on how you look at it: Scanpath comparison in multiple dimensions with multimatch, a vector-based approach. *Behavior Research Methods*, 44(4):1079–1100.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1):1–13.

Nora Hollenstein and Ce Zhang. 2019. Entity recognition at first sight: Improving NER with eye movement information. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1–10, Minneapolis, Minnesota. Association for Computational Linguistics.

Swati Jindal and Roberto Manduchi. 2023. Contrastive representation learning for gaze estimation. In *Gaze Meets Machine Learning Workshop*, pages 37–49. PMLR.

Varun Khurana, Yaman Kumar, Nora Hollenstein, Rajesh Kumar, and Balaji Krishnamurthy. 2023. Synthesizing human gaze feedback for improved NLP performance. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1895–1908, Dubrovnik, Croatia. Association for Computational Linguistics.

Walter Kintsch. 1992. A cognitive architecture for comprehension. In *Cognition: Conceptual and methodological issues*, pages 143–163. American Psychological Association, Washington, DC, US.

Walter Kintsch. 1994. Text comprehension, memory, and learning. *American Psychologist*, 49(4):294–303. Num Pages: 294-303 Place: Washington, US Publisher: American Psychological Association (US).

Bai Li and Frank Rudzicz. 2021. TorontoCL at CMCL 2021 shared task: RoBERTa with multi-stage fine-tuning for eye-tracking prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 85–89, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Abhijit Mishra, Kuntal Dey, and Pushpak Bhattacharyya. 2017. Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 377–387, Vancouver, Canada. Association for Computational Linguistics.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Alexander Pollatsek, Barbara J Juhasz, Erik D Reichle, Debra Machacek, and Keith Rayner. 2008. Immediate and delayed effects of word frequency and word length on eye movements in reading: a reversed delayed effect of word length. *Journal of experimental psychology: human perception and performance*, 34(3):726.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Erik D. Reichle, Keith Rayner, and Alexander Pollatsek. 2003. The E-Z reader model of eye-movement control in reading: Comparisons to other models. *The Behavioral and Brain Sciences*, 26(4):445–476; discussion 477–526.

Omer Shubi, David Robert Reich, Keren Gruteke Klein, Yuval Angel, Paul Prasse, Lena Ann Jäger, and Yevgeni Berzak. 2025. EyeBench: Predictive modeling from eye movements in reading. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ekta Sood, Prajit Dhar, Enrica Troiano, Rosy Southwell, and Sidney K. D'Mello. 2025. ScanEZ: Integrating cognitive models with self-supervised learning for spatiotemporal scanpath prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1132–1142, Vienna, Austria. Association for Computational Linguistics.

Ekta Sood, Simon Tannert, Philipp Mueller, and Andreas Bulling. 2020. Improving natural language processing tasks with human gaze-guided neural attention. In *Advances in Neural Information Processing Systems*, volume 33, pages 6327–6341. Curran Associates, Inc.

Harshvardhan Srivastava. 2022. Poirot at CMCL 2022 shared task: Zero shot crosslingual eye-tracking data prediction using multilingual transformer models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 102–107, Dublin, Ireland. Association for Computational Linguistics.

Ivan Stebakov and Ilya Pershin. 2025. How well can AI models generate human eye movements during reading? In *Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP)*, pages 148–162, Suzhou, China. Association for Computational Linguistics.

Haorui Sun, Ardyn Vivienne Olszko, Yida Chen, Hazen Kellner, Niharika Singh, Lincoln Lewisequerre, and David C. Jangraw. 2025. Reading observed at mindless moments (roamm): A simultaneous eeg and eye-tracking dataset of natural reading with attention annotations. In *NeurIPS 2025 Workshop on "Data on the Brain & Mind" Tutorials Track*.

Zihao Zhao, Sheng Wang, Qian Wang, and Dinggang Shen. 2024. Mining gaze for contrastive learning toward computer-assisted diagnosis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7):7543–7551.

A Appendix

A.1 Model Training

Below are details on our training pipeline.

Optimization. We train the model using the AdamW optimizer with a learning rate of 5×10^{-4}. We found that higher learning rates (e.g., 1×10^{-3}) led to gradient collapse and representation degeneration, where all embeddings converged to similar values. This learning rate is used consistently across all three training stages.

Batch Size. We use a batch size of 32 during contrastive training, providing 31 negative examples per positive pair. While larger batch sizes could provide more negative examples and potentially improve performance, we were constrained by GPU memory limitations. The resulting similarity matrix contains $N^2 = 1024$ comparisons per batch.

Data Preprocessing. Eye movement sequences are padded to the average length of time steps across datasets, with shorter sequences zero-padded and masked during attention. Text sequences are tokenized with the RoBERTa tokenizer, truncating longer texts and padding shorter ones.

Model Initialization. In Stage A, the language encoder is initialized from pretrained RoBERTa-base weights, while the ScanEZ encoder is initialized from the checkpoint of Sood et al. (2025). In Stage B, both encoders are initialized from their respective Stage A checkpoints. The contrastive projection heads introduced in this stage are randomly initialized. In Stage C, we discard the projection heads and fine-tune only the ScanEZ encoder and regression head.

Training Duration. Pretraining (Stage A) runs for 25 epochs, joint contrastive pretraining (Stage B) for 50 epochs, and scanpath prediction fine-tuning (Stage C) for 25 epochs. We monitor validation performance at each stage and save the best-performing checkpoint.

Training Datasets. We first fine-tune RoBERTa on the texts used to generate the synthetic scanpaths for ScanEZ pretraining with the E-Z Reader model (Reichle et al., 2003; Sood et al., 2025). Training is then continued on the respective ground-truth eye-tracking dataset of paired eye-movement–text examples, using the cross-validation procedure described in Section 5.

A.2 Evaluating Scanpath Prediction under Different Data Splits Conditions

The results in Section 6 are based on data splits in which neither participants nor textual stimuli overlap between training and test sets, preventing

Data	Model	$\text{NLL}_x\downarrow$	$\text{NLD}_x\downarrow$	$\text{NLL}_t\downarrow$	FDA↑	FLA↑
	ScanEZ	1.555 ±0.013	0.424 ±0.025	**1.235 ±0.072**	0.651 ±0.012	0.412 ± 0.007
CELER	**EYELIGNR**	**1.227 ± 0.010**	**0.166 ± 0.000**	1.464 ± 0.050	**0.688 ± 0.020**	**0.660 ± 0.000**
	ScanEZ	1.204 ± 0.021	0.336 ± 0.002	1.353 ± 0.033	0.714 ± 0.005	0.670 ± 0.006
ZuCo	**EYELIGNR**	**1.063 ± 0.014**	**0.328 ± 0.002**	**1.356 ± 0.034**	**0.724 ± 0.004**	**0.682 ± 0.007**
	ScanEZ	0.928 ± 0.037	0.585 ± 0.014	1.098 ± 0.026	0.803 ± 0.006	0.556 ± 0.013
EML	**EYELIGNR**	**0.835 ± 0.109**	**0.341 ± 0.002**	**1.018 ± 0.017**	**0.806 ± 0.006**	**0.735 ± 0.007**
	ScanEZ	2.894 ± 0.176	0.343 ± 0.004	2.274 ± 0.096	0.825 ± 0.001	0.833 ± 0.007
SB-SAT	**EYELIGNR**	**2.807 ± 0.190**	**0.335 ± 0.000**	**2.260 ± 0.090**	**0.829 ± 0.000**	**0.861 ± 0.010**
	ScanEZ	2.857 ± 0.125	0.344 ± 0.004	1.359 ± 0.094	0.841 ± 0.011	0.804 ± 0.004
ROAMM	**EYELIGNR**	**2.602 ± 0.118**	**0.303 ± 0.004**	**1.323 ± 0.095**	**0.848 ± 0.009**	**0.817 ± 0.004**

Table 4: Scanpath prediction results on the participant-disjoint train/test splits.

Data	Model	$\text{NLL}_x\downarrow$	$\text{NLD}_x\downarrow$	$\text{NLL}_t\downarrow$	FDA↑	FLA↑
	ScanEZ	1.603 ± 0.017	0.430 ± 0.047	**1.269** ± 0.068	0.653 ± 0.011	0.408 ± 0.008
CELER	**EYELIGNR**	**1.418 ± 0.044**	**0.276 ± 0.001**	1.609 ± 0.032	**0.678 ± 0.004**	**0.518 ± 0.006**
	ScanEZ	1.045 ± 0.031	0.329 ± 0.003	1.479 ± 0.036	0.698 ± 0.004	0.665 ± 0.001
ZuCo	EYELIGNR	**1.060 ± 0.014**	**0.319 ± 0.004**	**1.296 ± 0.035**	**0.739 ± 0.004**	**0.690 ± 0.006**
	ScanEZ	1.000 ± 0.012	0.589 ± 0.012	1.106 ± 0.022	0.803 ± 0.004	0.553 ± 0.020
EML	**EYELIGNR**	**0.791 ± 0.039**	**0.335 ± 0.001**	**1.084 ± 0.005**	**0.813 ± 0.003**	**0.733 ± 0.003**
	ScanEZ	3.052 ± 0.198	0.353 ± 0.004	2.394 ± 0.058	0.819 ± 0.009	0.807 ± 0.005
SB-SAT	**EYELIGNR**	**2.908 ± 0.018**	**0.345 ± 0.003**	**2.299 ± 0.073**	**0.828 ± 0.008**	**0.818 ± 0.004**
	ScanEZ	2.717 ± 0.047	0.357 ± 0.002	**1.336 ± 0.019**	0.836 ± 0.002	0.801 ± 0.002
ROAMM	**EYELIGNR**	**2.555 ± 0.045**	**0.349 ± 0.001**	1.371 ± 0.021	**0.842 ± 0.002**	**0.810 ± 0.001**

Table 5: Scanpath prediction results on the textual stimulus-disjoint train/test splits.

models from exploiting document- or participant-specific regularities.

Here, we experiment with more lenient types of data splits to enable a full comparison with the evaluation protocol of Sood et al. (2025). Specifically, we consider two strategies that enforce no overlap between training and test sets with respect to participants or textual stimuli, respectively.

The results (Tables 4 and 5) align with our earlier findings: with the exception of NLL_t on CELER, EYELIGNR consistently outperforms ScanEZ.

A.3 Pointwise and Sequence-Level Metrics

Table 6 reports the full statistical comparisons of EYELIGNR with ScanEZ and Token-Ordered, including paired significance tests and effect sizes. These comparisons are based on model performance measured using RMSE to quantify how closely the predicted trajectories match the ground truth. We also report Explained Variance (EV). For CELER, where scanpaths have no vertical variation due to the stimuli being single-line sentences, EV in y is calculated using a fallback formulation with prediction variance in the denominator when $\text{Var}(\text{y}) = 0$, to avoid undefined values. Tests are conducted separately for each dimension. Except for the vertical dimension (y), we observe that EYELIGNR exhibits significantly lower RMSE and higher EV than the baseline models.

The RMSE and EV evaluation metrics focus on *local, pointwise* errors. To complement them with a more holistic sequence-level perspective, we investigate the quality of predictions using MultiMatch (Dewhurst et al., 2012). Results (Table 7) corroborate and refine our earlier insights. We see that both EYELIGNR and ScanEZ achieve high similarity in coarse spatial structure (shape and length). However, EYELIGNR yields substantially higher directional similarity across datasets, indicating more human-like patterns of saccades and regressions. This improvement in directionality is consistent with EYELIGNR's stronger modeling of within-line reading dynamics, which are closely tied to linguistic processing, and reflects sequence-level gains that are not captured by pointwise error metrics alone. The comparison between EYELIGNR and the Token-Ordered baseline is discussed in the body of the paper.

			RMSE				EV			
Data	**Dim.**	**Model**	$\mathbf{p_{adj}}$	**d**	**Magn.**	**Best**	$\mathbf{p_{adj}}$	**d**	**Magn.**	**Best**
CELER	x	ScanEZ	< .001	-0.196	negl.	EyeLignR	< .001	0.693	med.	EyeLignR
		Token-Ordered	< .001	-2.310	large	EyeLignR	< .001	0.749	med.	EyeLignR
	y	ScanEZ	0.126	0.091	negl.	ScanEZ	1.000	-0.042	negl.	ScanEZ
		Token-Ordered	< .001	-1.628	large	EyeLignR	1.000	-0.043	negl.	Token-Ordered
	t	ScanEZ	< .001	0.246	small	ScanEZ	< .001	-0.345	small	ScanEZ
		Token-Ordered	< .001	-1.453	large	EyeLignR	< .001	0.378	small	EyeLignR
ZuCo	x	ScanEZ	< .001	-0.837	large	EyeLignR	< .001	1.351	large	EyeLignR
		Token-Ordered	< .001	-3.701	large	EyeLignR	< .001	3.132	large	EyeLignR
	y	ScanEZ	< .001	1.021	large	ScanEZ	< .001	-0.536	med.	ScanEZ
		Token-Ordered	< .001	-5.160	large	EyeLignR	< .001	1.580	large	EyeLignR
	t	ScanEZ	< .001	0.558	med.	ScanEZ	1.000	0.111	negl.	EyeLignR
		Token-Ordered	< .001	-1.938	large	EyeLignR	< .001	2.030	large	EyeLignR
EML	x	ScanEZ	< .001	-1.484	large	EyeLignR	< .001	1.357	large	EyeLignR
		Token-Ordered	< .001	-1.868	large	EyeLignR	< .001	2.982	large	EyeLignR
	y	ScanEZ	< .001	0.547	med.	ScanEZ	0.006	-0.143	negl.	ScanEZ
		Token-Ordered	< .001	-3.339	large	EyeLignR	< .001	2.509	large	EyeLignR
	t	ScanEZ	1.000	-0.057	negl.	EyeLignR	0.085	0.116	negl.	EyeLignR
		Token-Ordered	< .001	-1.336	large	EyeLignR	< .001	2.119	large	EyeLignR
ROAMM	x	ScanEZ	< .001	-1.545	large	EyeLignR	< .001	1.354	large	EyeLignR
		Token-Ordered	< .001	-4.964	large	EyeLignR	< .001	5.397	large	EyeLignR
	y	ScanEZ	0.166	0.196	negl.	ScanEZ	0.034	-0.228	small	ScanEZ
		Token-Ordered	< .001	-5.477	large	EyeLignR	< .001	1.570	large	EyeLignR
	t	ScanEZ	0.015	-0.243	small	EyeLignR	0.213	0.190	negl.	EyeLignR
		Token-Ordered	< .001	-1.980	large	EyeLignR	< .001	1.499	large	EyeLignR
SB-SAT	x	ScanEZ	< .001	-2.917	large	EyeLignR	< .001	2.530	large	EyeLignR
		Token-Ordered	< .001	-6.378	large	EyeLignR	< .001	7.048	large	EyeLignR
	y	ScanEZ	< .001	-0.814	large	EyeLignR	< .001	0.897	large	EyeLignR
		Token-Ordered	< .001	-5.781	large	EyeLignR	< .001	7.902	large	EyeLignR
	t	ScanEZ	1.000	-0.076	negl.	EyeLignR	1.000	0.079	negl.	EyeLignR
		Token-Ordered	< .001	-2.761	large	EyeLignR	< .001	2.783	large	EyeLignR

Table 6: Paired comparisons of the baseline models with EyeLignR with respect to RMSE and EV across the normalized dimensions x, y, and t (Dim.). Bonferroni-corrected p-values and paired-sample Cohen's d are rounded to three decimals. Best: the model which predicts most similar to the ground truth. Mag.: magnitude of the effect size. Med. and negl.: medium and negligible effect sizes, respectively.

Data	Model	Shape	Dir	Len	Pos	Dur
CELER	ScanEZ	0.986 ± 0.00	0.717 ± 0.01	0.985 ± 0.01	0.928 ± 0.00	0.772 ± 0.01
	Token-Ordered	0.986 ± 0.02	0.652 ± 0.01	0.984 ± 0.01	0.827 ± 0.03	0.579 ± 0.01
	EYELIGNR	**0.993 ± 0.00**	**0.977 ± 0.00**	**0.988 ± 0.00**	**0.971 ± 0.00**	**0.787 ± 0.01**
ZuCo	ScanEZ	0.979 ± 0.00	0.877 ± 0.01	0.969 ± 0.01	0.923 ± 0.00	0.769 ± 0.01
	Token-Ordered	0.976 ± 0.00	0.778 ± 0.01	0.971 ± 0.00	0.665 ± 0.01	0.645 ± 0.01
	EYELIGNR	**0.980 ± 0.00**	**0.917 ± 0.00**	**0.972 ± 0.00**	**0.929 ± 0.00**	**0.783 ± 0.01**
EML	ScanEZ	0.982 ± 0.00	0.901 ± 0.00	0.976 ± 0.00	0.935 ± 0.00	0.784 ± 0.02
	Token-Ordered	0.973 ± 0.00	0.704 ± 0.00	0.973 ± 0.00	0.818 ± 0.03	0.642 ± 0.01
	EYELIGNR	**0.983 ± 0.00**	**0.908 ± 0.02**	**0.977 ± 0.00**	**0.938 ± 0.01**	**0.789 ± 0.04**
SB-SAT	ScanEZ	0.973 ± 0.00	0.835 ± 0.01	0.967 ± 0.00	0.887 ± 0.00	0.728 ± 0.01
	Token-Ordered	0.971 ± 0.00	0.740 ± 0.01	0.969 ± 0.00	0.742 ± 0.01	0.380 ± 0.02
	EYELIGNR	**0.979 ± 0.00**	**0.915 ± 0.00**	**0.971 ± 0.00**	**0.922 ± 0.00**	**0.741 ± 0.01**
ROAMM	ScanEZ	0.976 ± 0.00	0.853 ± 0.01	0.970 ± 0.01	0.901 ± 0.03	0.765 ± 0.01
	Token-Ordered	0.972 ± 0.00	0.740 ± 0.00	0.969 ± 0.00	0.734 ± 0.02	0.499 ± 0.02
	EYELIGNR	**0.979 ± 0.00**	**0.912 ± 0.00**	**0.972 ± 0.00**	**0.925 ± 0.00**	**0.779 ± 0.00**

Table 7: MultiMatch results (mean and standard deviation) for all used models. For all metrics, the higher the better.