ScanEZ: Integrating Cognitive Models with Self-Supervised Learning for Spatiotemporal Scanpath Prediction

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

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

Enrica Troiano
HK3Lab
enrica.troiano@hk3lab.ai

Rosy Southwell
University of Colorado Boulder
rosy.southwell@colorado.edu

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

Abstract

Accurately predicting human scanpaths during reading is vital for diverse fields and downstream tasks, from educational technologies to automatic question answering. To date, however, progress in this direction remains limited by scarce gaze data. We overcome the issue with ScanEZ, a self-supervised framework grounded in cognitive models of reading. ScanEZ jointly models the spatial and temporal dimensions of scanpaths by leveraging synthetic data and a 3-D gaze objective inspired by masked language modeling. With this framework, we provide evidence that two key factors in scanpath prediction during reading are: the use of masked modeling of both spatial and temporal patterns of eye movements, and cognitive model simulations as an inductive bias to kick-start training. Our approach achieves state-of-the-art results on established datasets (e.g., up to 31.4% negative log-likelihood improvement on CELER L1), and proves portable across different experimental conditions.

1 Introduction

Research at the intersection of cognitive science, computer vision (CV), and natural language processing (NLP) shows the potential for modeling human reading comprehension through eye-tracking data. Many downstream tasks, like entity recognition (Hollenstein and Zhang, 2019), summarization (Stiennon et al., 2020; Wu et al., 2021), and question answering (Sood et al., 2023), have particularly benefited from the use of scanpaths (i.e., records of eye movements on text), which provide signal about the readers' cognitive processes (Rayner, 1998) as well as the features of the text being read (Scozzaro et al., 2024; Wiechmann et al., 2022).

Scanpath datasets, however, are hard to obtain, and their scarcity represents a bottleneck for training gaze-aware architectures (Kümmerer et al., 2016). To solve this problem, past research has taken either of two directions: exploring self-supervised learning (SSL) techniques to predict eye gaze using limited labeled samples (Islam et al., 2021, i.a.), or augmenting existing samples with synthetic ones (Khurana et al., 2023). Works based on these approaches circumvented data shortages, but analyzed scanpath patterns to a limited extent: they modeled spatial dynamics (i.e., the order of words being fixated) that implicitly contain the chronology of fixations; yet, they overlooked *fixation duration* (i.e., how long the eyes dwell on a word). Duration is a key temporal aspect of reading because it links to word processing difficulty (Rayner, 1998; Clifton Jr et al., 2007; Vasishth et al., 2012). Nevertheless, related NLP research hardly incorporates this and other core linguistic and physiological factors that are well captured by cognitive models of reading behavior (Reichle et al., 2003; Engbert et al., 2005; Salvucci, 2001) – e.g., the effects of syntax complexity on word fixation (Cook and Wei, 2019), and the limits of visual range on eye movement (McDonald and Shillcock, 2003; McNamara and Magliano, 2009).

In this paper, we fill these gaps by combining data augmentation and SSL techniques: we propose ScanEZ, a SSL framework that uses synthetic data (grounded in cognitive models of reading) to predict gaze trajectories in terms of both spatial and temporal aspects, including fixation duration. With respect to data augmentation, ScanEZ uses the E-Z Reader cognitive model (Reichle et al., 2003) to derive synthetic gaze samples from existing corpora. On the modeling side, the framework jointly learns the spatial (x, y) and temporal (t) coordinates of gaze by employing a masked gaze modeling objective. Masked language modeling, a well-established approach in NLP and CV (Devlin et al., 2019; Zhang et al., 2022; Kwon et al., 2022), is effective for capturing rich sequence dependencies. We demonstrate its utility to predict full 3-D spatiotemporal trajectories of scanpaths, which inherently rely on textual dependencies. Ex-

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 1132–1142

perimental results demonstrate strong performance across datasets, including CELER L1 (Berzak et al., 2022), where ScanEZ surpasses the state of the art (Deng et al., 2023) by up to 31.4% (negative log-likelihood), and ZuCo 1.0 (Hollenstein, 2018), with a 58.6% improvement. We also evaluate our approach with fixation duration and fixation location accuracy, two eye-movement metrics which assess the model's ability to capture *how long* the eyes dwell on a word, reflecting temporal attention in reading, and *where* the eyes fixate, reflecting spatial attention.

In sum, this work advances scanpath prediction with the following contributions: (1) a novel framework to model spatial and temporal dimensions simultaneously, integrating cognitive data with scalable NLP techniques (notably, masked language modeling); (2) state-of-the-art scanpath prediction results, complemented by analyses to assess cross-domain generalization performance, and the role of synthetic data pre-training and human-data fine-tuning; (3) additional model evaluation based on eye-tracking metrics that capture the interaction between text and readers.

2 Related Work

Reading comprehension involves multiple cognitive processes (Kintsch, 1998), from word and sentence processing to integrating the text with a reader's prior knowledge (Kintsch, 1994; McNamara and Magliano, 2009). These processes unfold chronologically, but they extend beyond simple left-to-right reading, as eye movements have non-linear patterns of word fixation and skipping (Engbert et al., 2004). Accurately modeling both spatial and temporal aspects of eye movements is thus essential to understand reading behavior.

Empirical evidence is plentiful for the link between specific eye movement features (fixation duration, location, etc.) and key linguistic and cognitive phenomena, for which specific measures have been proposed (Cook and Wei, 2019; McDonald and Shillcock, 2003; Shain, 2024; Just and Carpenter, 1976; Rayner et al., 1998). These features are also captured by computational cognitive models which render explicit the relationship between cognitive processes, like word recognition, and physical actions, such as word fixations. Among these, E-Z Reader (Reichle et al., 2003) simulates the control of eye movements on a given text. E-Z Reader has been shown to approximate human gaze behavior across a range of datasets and experimental conditions (Mancheva et al., 2015; Reichle and Drieghe, 2013; Reichle and Sheridan, 2015, i.a.,). Its predictive validity has been further supported by evaluations against human eye-tracking data, including comparisons on the CELER dataset (Deng et al., 2023). Moreover, E-Z Reader-based simulations over the CNN and DM corpora have demonstrated utility in downstream tasks, achieving top performance in text saliency prediction when modeled during pre-training (Sood et al., 2020).

While replicating core reading patterns, E-Z Reader relies on handcrafted features that limit its scalability for AI applications. Deep learning offers alternatives which can closely approximate human reading patterns. For instance, SCANDL uses a sequence-to-sequence diffusion model to capture gaze-text interactions (Bolliger et al., 2023), SP-EyeGAN simulates raw eye-tracking data with generative adversarial networks (Prasse et al., 2023), and Eyettention (Deng et al., 2023) introduces a dual-sequence encoder-decoder architecture with cross-attention mechanisms to align linguistic and temporal sequences to predict the next fixation location. SCANDL and SP-EyeGAN further address the paucity of eye-tracking resources with their ability to generate synthetic data. In part, the lack of data is also mitigated by SSL (Ericsson et al., 2022) approaches, which derive gaze features from unlabeled samples. Contrastive and predictive objectives, for instance, have been used to capture statistical regularities in fixation patterns to support downstream tasks (Prasse et al., 2024). However, SSL still requires large-scale data.

In this paper, we adopt SSL like SCANDL and SP-EyeGAN, while pre-training on fixations generated with the E-Z Reader cognitive model. This way, we push masked language modeling for reading gaze representation to be data-rich and cognitively-driven (Sood et al., 2020). Notably, we extend the approach of Deng et al. (2023), who model spatial scanpath coordinates for scanpath prediction, by simultaneously incorporating temporal coordinates (fixation durations). We directly compare our results to theirs, as Eyettention represents the current state of the art. To enrich model evaluation, we further employ experimental eye-movement metrics. As part of this framework, we use synthetic corpora introduced in prior work, relying on the established validity of the E-Z Reader model as a synthetic data generator.

3 ScanEZ

We propose a framework that adapts a BERT-style transformer for trajectory modeling, using a masked sequence prediction objective and synthetic data pre-training to enable representation learning without large labeled datasets.

Input Representation and Masking. Inputs to the model are gaze data parsed into sequences of fixations, each represented by three features describing the trajectory of gaze over text: the x- and y-coordinates (space) and fixation duration t (time). For preprocessing, each input sequence $X \in \mathbb{R}^{T \times F}$, where T denotes the number of time steps and $F = 3$ the number of features, is normalized to zero mean and unit variance. Normalization ensures numerical stability during training and harmonizes the distributions of real and synthetic data.

During pre-training, a subset of input features is masked: a fixed proportion of the sequence is obscured, generating X_{masked}, where partial information is visible to the model. In the prediction task, the model predicts the masked values (x, y, t) based on surrounding context, learning dependencies across the spatial and temporal dimensions.

Model Architecture. The network consists of an embedding layer which projects the input features (x, y, t) into a latent space of dimension d, creating dense embeddings that preserve spatiotemporal relationships. Sinusoidal positional encodings are added to the embeddings, ensuring that the model distinguishes between otherwise position-invariant elements in the sequence. The model also comprises L transformer layers with a multi-head self-attention mechanism (h attention heads attend to relationships across time steps and between spatial and temporal features), and a fully connected feedforward network with two linear layers and ReLU activations (which capture interactions between features).

Training Objective. The pre-training task is to reconstruct masked portions of the input sequence. Given a masked input X_{masked}, the model predicts the values of the masked features (x, y, t):

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

where $\mathcal{M}$ denotes the set of masked indices, X_i the ground truth, and $\hat{X}_i$ the predicted values. This loss encourages the model to learn embeddings that capture both local and global spatiotemporal dependencies in sequential gaze data.

We pre-train the model on synthetic 3-D trajectories, and fine-tune its weights on real eye-tracking datasets – both sources of data are discussed next.

4 Experimental Setup

We now describe the datasets and evaluation protocols used for scanpath prediction. To better reflect real-world reading scenarios, we selected diverse benchmark datasets and adopted multiple evaluation settings, including three distinct data-splitting approaches. Further details on the data and evaluation metrics are in appendix A.1.

Data. As illustrated in Figure 1, we pre-train ScanEZ on E-Z Reader simulations, which were generated by Sood et al. (2020) from the CNN and Daily Mail corpus (Hermann et al., 2015). For fine-tuning, we use two human sentence-reading datasets with the same train/validation splits for the scanpath prediction task as in Deng et al. (2023). These datasets vary in complexity, domain, and participant diversity, enabling in-domain and cross-domain robustness tests: *CELER L1* (Berzak et al., 2022), with native (L1) English speakers reading single-line sentences from Wall Street Journal (Marcus et al., 1993): and *ZuCo 1.0* (Hollenstein et al., 2018), with L1 English speakers reading sentences from the Stanford Sentiment Treebank (Socher et al., 2013) and Wikipedia relation extraction corpus (Culotta et al., 2006), with sentences displayed as multiple (maximum 7) lines. In addition, we use *EML* (Caruso et al., 2022), a dataset of fluent English speakers reading 5 complex educational texts, each of ≈1000 words, displayed as 50 full pages with ≈100 words on each, from which we extract 239 sentences.

Evaluation. We analyze performance in the four cross-validation task settings reported in Figure 1. *Participant split* (Part): Train/validate on distinct sets of participants. *Text split* (Text): Train/validate on distinct texts (i.e. sentences or pages)[1]. *Participant + Text split* (P.T.): No overlap in participants nor texts between training and validation. This is the most challenging condition for the within dataset evaluation settings (Deng et al., 2023). *Cross-dataset*: Train on a dataset and test on another, to assess cross-domain generalization. We keep the same 5-fold evaluation protocol across settings for robust performance estimates.

[1]We use *text* for any stimulus shown to participants. Sentence count per stimulus varies by dataset: Celer ≈ 1; ZuCo ≈ 2–3; EML ≈ 5.

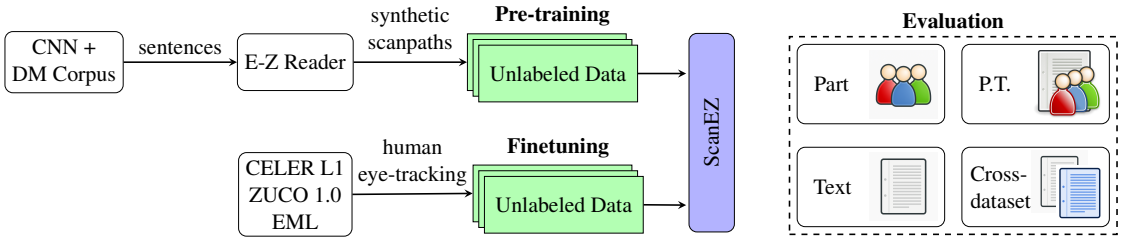

Figure 1: Overview of the workflow combining synthetic and human eye-tracking data for scan-path prediction. Synthetic scanpaths generated with the E-Z Reader model from CNN + DM texts are used in the pre-training phase of ScanEZ. The model is then fine-tuned on real human data (CELER, ZuCo 1.0, EML). We evaluate ScanEZ's performance in four cross-validation settings – Part(), Text(), P.T.(), and Cross-dataset() – described Section 4.

For spatial predictions, we report Negative Log-Likelihood (NLL), to measure how well the model fits observed scan paths, and Normalized Levenshtein Distance (NLD), quantifying alignment quality between predicted and actual sequences. For temporal predictions, we use NLL_t, as NLD considers the distance between spatial sequences (e.g., strings), which does not apply to the temporal dimension. To better observe eye movement characteristics, we also include fixation location accuracy (FLA) and fixation duration accuracy (FDA), which range from 0 (worst possible accuracy) to 1 (best).

5 Does ScanEZ Capture Where Readers Fixate, and When They Do So?

We answer this question by replicating the experiments of Deng et al. (2023) and observing ScanEZ performance on naturalistic (EML) data.[2]

Comparison with Eyettention. ScanEZ achieves better (i.e., lower) NLL and NLD under all experimental settings. For instance, on the Text split of CELER L1, it has a NLL of 1.603 compared to the 2.277 of Eyettention, and a NLD of 0.43, which is 0.142 points more than Eyettention; on the Part split, it brings an improvement of 0.712 points in NLL (NLL = 1.555, against Eyettention's 2.267) and of 0.149 points in NLD (NLD = 0.424 for ScanEZ, 0.573 for Eyettention). Even on P.T., which is our most challenging condition, ScanEZ surpasses Eyettention on CELER L1. Table 1 focuses on this setting, and shows that ScanEZ performs better by 0.77 and 0.147 points, with a NLL score of 1.524 and a NLD score of 4.421.

This improvement remains valid in the cross-dataset setup: on the P.T. split, our framework achieves NLL = 0.548 when trained on CELER L1 and tested on ZuCo 1.0, whereas Eyettention has a score of 2.613. On average, the P.T. NLL is reduced by 58.6% compared to the baseline when testing on ZuCo 1.0 across train-testing combinations.[3] These scores corroborate the value of masked-language modeling in scanpath prediction, promoting out-of-domain generalization.

Table 1 also provides evidence for the benefit of pre-training and fine-tuning ScanEZ: removing human-based fine-tuning brings NLL up to 3.035, which underperforms both ScanEZ and Eyettention, and models without pre-training are better than Eyettention (potentially thanks to the fine-tuning step) but still worse than ScanEZ (1.77 NLL). Similar results hold for the other metrics as well: ScanEZ outperforms its ablation alternatives with respect to FDA and FLA; moreover, it sets benchmark performance for the temporal prediction in terms of NLL_t.[4] Overall, fine-tuning enables the model to efficiently capture eye movement patterns in real eye-tracking data, and kick-starting training with E-Z Reader simulations supports robust initial representations. Put together, these components allow ScanEZ to deliver superior performance in spatial scanpath prediction, and efficiency in the modeling of its temporal dimension.

Evaluations on EML We analyze the performance of ScanEZ and its ablated variants on the EML dataset (Caruso et al., 2022). Once more, pre-training yields pronounced improvements across metrics (Table 1, bottom). An exception is FLA,

[2]We report direct comparisons only with Deng et al. (2023), as their work includes extensive evaluations against a broad range of models. This allows to interpret our findings in the light of those prior comparisons.

[3]A full by-setting breakdown of all results discussed in this section is in appendix A.2 (Table 3), including performance on train-test set combinations (Table 4), and ablation analyses to isolate the role of pre-training and fine-tuning (Table 5).

[4]These three metrics are unavailable for Eyettention.

Data	Model	NLL↓	NLD↓	NLL$_t$↓	FDA↑	FLA↑
CLR	Eyettention	2.297 ± 0.011	0.568 ± 0.004	—	—	—
	w/o Fine-tuning	3.035 ± 0.204	0.950 ± 0.009	1.403 ± 0.012	0.603 ± 0.004	0.284 ± 0.018
	w/o Pre-training	1.772 ± 0.304	0.547 ± 0.085	1.431 ± 0.150	0.611 ± 0.054	0.381 ± 0.054
	ScanEZ	**1.524 ± 0.042**	**0.421 ± 0.022**	**1.244 ± 0.027**	**0.646 ± 0.004**	**0.413 ± 0.009**
EML	w/o Fine-tuning	1.837± 0.025	0.761± 0.005	1.181± 0.011	0.797± 0.003	**0.589± 0.006**
	w/o Pre-training	1.217± 0.212	0.777± 0.004	1.177± 0.066	0.785± 0.012	0.491± 0.023
	ScanEZ	**0.996± 0.020**	**0.616± 0.013**	**1.083± 0.033**	**0.804± 0.005**	0.534± 0.012

Table 1: Top: comparison between our framework and Eyettention on the CELER L1 dataset (CLR) to our model trained on: only EZ-Reader data (w/o Fine-tuning), only human data (w/o Pre-training), and with both pre-training on EZ-Reader and then fune-tuning on human data (ScanEZ). Bottom: evaluation using the EML dataset. ↓: the lower the score, the better. ↑: vice versa. All results refer to the P.T. setting.

significantly higher for the w/o Fine-tuning setting than for ScanEZ (p-value = 0.001, with an ANOVA test). This result is inconsistent with that on CELER L1: we attribute it to EML's more challenging data, which comprises multi-sentences texts rather than single sentence texts, making the task more complex.[5] While future work could investigate this insight, the results presented here maintain that ScanEZ effectively generalizes to various reading conditions.

6 Conclusion

We introduced ScanEZ, a self-supervised framework that addresses the challenge of scan path prediction under data scarcity conditions. Aimed at an explicit spatiotemporal modeling of eye movements, ScanEZ demonstrated robust performance and generalization abilities across datasets, experimental settings, and metrics, surpassing the state of the art on spatial predictions and benchmarking the prediction of fixation duration. Importantly, our results point to the benefits of the masked-language modeling approach we implemented, which opens up new research directions on eye-tracking data. As a matter of fact, since masked language modeling has not been as widely adopted for scanpath prediction as they have in NLP tasks, our finding that a BERT-based model yields superior performance constitutes a core contribution of this work. Our ablation analyses further showed the importance of real-data fine-tuning for a model pre-trained on synthetic gaze trajectories. Both components, which underpin the value of bridging cognitive model simulations with data-driven methodologies for gaze representation, allowed ScanEZ to effectively capture spatiotemporal dependencies.

[5]These FLA differences are analyzed in appendix A.2.

7 Limitations

We evaluated ScanEZ on two established open-source datasets (CELER L1 and Zuco 1.0). This experimental decision ensured comparability with previous work, specifically with Eyettention. However, to fully understand the generalizability of the model we proposed, we would need to conduct additional evaluations, in particular on datasets representing real-world reading scenarios – unlike CELER L1 and ZuCo 1.0, which were built in a lab environment. We took a step in this direction by reporting results on the EML data (drawn from a naturalistic e-learning setting). Yet, EML is not open source, which might complicate the reproduction of our results for other researchers.

Focusing on such results, we reported ScanEZ's success in modeling both spatial and temporal aspects of gaze. The comparison with past work, however, was only partial, as NLLt, fixation duration accuracy (FDA), and fixation location accuracy (FLA) are not reported for Eyettention.

Lastly, by addressing scanpath prediction, our work contributes to an established line of research that studies eye movements (e.g., Boccignone et al., 2019; de Belen et al., 2022; Mishra et al., 2018). As such, it represents a step towards evaluating gaze-informed models in applications related to reading behavior. To date, however, we still need to test ScanEZ on downstream tasks, to better understand the breadth of its potential.

8 Ethical Considerations

This work did not go through an ethical committee; still, we ensured compliance with ethical principles in data usage and methodology. On the one hand, we used artificial gaze trajectories – no real individuals can be reconstructed from it. On the other,

we relied on publicly available datasets and EML (all already peer-reviewed) where 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

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.

Giuseppe Boccignone, Vittorio Cuculo, and Alessandro D'Amelio. 2019. How to look next? a data-driven approach for scanpath prediction. In *International Symposium on Formal Methods*, pages 131–145. Springer.

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.

Megan Caruso, Candace Peacock, Rosy Southwell, Guojing Zhou, and Sidney D'Mello. 2022. Going Deep and Far: Gaze-based Models Predict Multiple Depths of Comprehension During and One Week Following Reading.

Charles Clifton Jr, Adrian Staub, and Keith Rayner. 2007. Eye movements in reading words and sentences. *Eye movements*, pages 341–371.

Anne E. Cook and Wei Wei. 2019. What Can Eye Movements Tell Us about Higher Level Comprehension? *Vision*, 3(3):45. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

Aron Culotta, Andrew McCallum, and Jonathan Betz. 2006. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 296–303, New York City, USA. Association for Computational Linguistics.

Ryan Anthony Jalova de Belen, Tomasz Bednarz, and Arcot Sowmya. 2022. Scanpathnet: A recurrent mixture density network for scanpath prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5010–5020.

Shuwen Deng, David R. Reich, Paul Prasse, Patrick Haller, Tobias Scheffer, and Lena A. Jäger. 2023. Eyettention: An Attention-based Dual-Sequence Model for Predicting Human Scanpaths during Reading. ArXiv:2304.10784 [cs].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Ralf Engbert, Reinhold Kliegl, and André Longtin. 2004. Complexity of eye movements in reading. *International Journal of Bifurcation and Chaos*, 14(02):493–503.

Ralf Engbert, Antje Nuthmann, Eike M. Richter, and Reinhold Kliegl. 2005. SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, 112(4):777–813.

Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M. Hospedales. 2022. Self-Supervised Representation Learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3):42–62.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Nora Hollenstein. 2018. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*.

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):180291. Number: 1 Publisher: Nature Publishing Group.

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.

Md. Rabiul Islam, Shuji Sakamoto, Yoshihiro Yamada, Andrew W. Vargo, Motoi Iwata, Masakazu Iwamura, and Koichi Kise. 2021. Self-supervised learning for reading activity classification. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5(3).

Marcel Adam Just and Patricia A. Carpenter. 1976. Eye fixations and cognitive processes. *Cognitive Psychology*, 8(4):441–480.

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

Walter Kintsch. 1998. *Comprehension: A Paradigm for Cognition*. Cambridge University Press.

Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. 2016. Deepgaze II: reading fixations from deep features trained on object recognition. *CoRR*, abs/1610.01563.

Gukyeong Kwon, Zhaowei Cai, Avinash Ravichandran, Erhan Bas, Rahul Bhotika, and Stefano Soatto. 2022. Masked vision and language modeling for multi-modal representation learning. *arXiv preprint arXiv:2208.02131*.

Lyuba Mancheva, Erik D. Reichle, Benoît Lemaire, Sylviane Valdois, Jean Ecalle, and Anne Guérin-Dugué. 2015. An analysis of reading skill development using e-z reader. *Journal of Cognitive Psychology (Hove)*, 27(5):357–373. Epub 2015 Apr 9. PMID: 27148437, PMC4852752.

Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330.

Scott A. McDonald and Richard C. Shillcock. 2003. Eye Movements Reveal the On-Line Computation of Lexical Probabilities During Reading. *Psychological Science*, 14(6):648–652. Publisher: SAGE Publications Inc.

Danielle S. McNamara and Joe Magliano. 2009. Chapter 9 Toward a Comprehensive Model of Comprehension. In *Psychology of Learning and Motivation*, volume 51 of *The Psychology of Learning and Motivation*, pages 297–384. Academic Press.

Abhijit Mishra, Pushpak Bhattacharyya, Abhijit Mishra, and Pushpak Bhattacharyya. 2018. Scanpath complexity: modeling reading/annotation effort using gaze information. *Cognitively Inspired Natural Language Processing: An Investigation Based on Eye-tracking*, pages 77–98.

Paul Prasse, David R. Reich, Silvia Makowski, Tobias Scheffer, and Lena A. Jäger. 2024. Improving cognitive-state analysis from eye gaze with synthetic eye-movement data. *Computers & Graphics*, 119:103901.

Paul Prasse, David Robert Reich, Silvia Makowski, Seoyoung Ahn, Tobias Scheffer, and Lena A. Jäger. 2023. Sp-eyegan: Generating synthetic eye movement data with generative adversarial networks. In *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*, ETRA '23, New York, NY, USA. Association for Computing Machinery.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372.

Keith Rayner, Erik D. Reichle, and Alexander Pollatsek. 1998. Eye Movement Control in Reading: An Overview and Model. *Eye guidance in reading and scene perception*, pages 243–268.

Erik D. Reichle and Denis Drieghe. 2013. Using EZ reader to examine word skipping during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(4):1311.

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.

Erik D. Reichle and Heather Sheridan. 2015. EZ reader: An overview of the model and two recent applications. *The Oxford handbook of reading*, page 277.

Dario D. Salvucci. 2001. An integrated model of eye movements and visual encoding. *Cognitive Systems Research*, 1(4):201–220.

Calogero J. Scozzaro, Davide Colla, Matteo Delsanto, Antonio Mastropaolo, Enrico Mensa, Luisa Revelli, and Daniele P. Radicioni. 2024. Legal text reader profiling: Evidences from eye tracking and surprisal based analysis. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, pages 114–124, Torino, Italia. ELRA and ICCL.

Cory Shain. 2024. Word Frequency and Predictability Dissociate in Naturalistic Reading. *Open Mind*, 8:177–201.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Ekta Sood, Fabian Kögel, Philipp Müller, Dominike Thomas, Mihai Bâce, and Andreas Bulling. 2023. Multimodal Integration of Human-Like Attention in Visual Question Answering. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2648–2658, Vancouver, BC, Canada. IEEE.

Ekta Sood, Simon Tannert, Philipp Mueller, and Andreas Bulling. 2020. Improving Natural Language Processing Tasks with Human Gaze-Guided Neural Attention.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Shravan Vasishth, Titus von der Malsburg, and Felix Engelmann. 2012. What can eye-tracking tell us about sentence processing. *WIREs Cognitive Science*.

Daniel Wiechmann, Yu Qiao, Elma Kerz, and Justus Mattern. 2022. Measuring the impact of (psycho-)linguistic and readability features and their spill over effects on the prediction of eye movement patterns. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5276–5290, Dublin, Ireland. Association for Computational Linguistics.

Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*.

Chaoning Zhang, Chenshuang Zhang, Junha Song, John Seon Keun Yi, Kang Zhang, and In So Kweon. 2022. A survey on masked autoencoder for self-supervised learning in vision and beyond. *arXiv preprint arXiv:2208.00173*.

A Appendix

A.1 Details on Experimental Setup

Data. Table 2 reports statistics on the data used to pre-train and fine-tune ScanEZ.

Pre-training	Sentences	Simulations
E-Z Reader: CNN	3.1M	10
E-Z Reader: Daily Mail	7.6M	10
Fine-tuning	Sentences	Participants
CELER L1	5457	69
ZuCo 1.0	700	12
EML	239	147

Table 2: Descriptive statistics of the used datasets. Top: synthetic datasets used for pre-training; bottom: human datasets used in our experiments.

Metrics. The two metrics FDA (Fixation Duration Accuracy) and FLA (Fixation Location Accuracy) are computed as follows:

$$\text{FDA} = 1 - \frac{|T_{\text{pred}} - T_{\text{ground}}|}{\max(|T_{\text{pred}} - T_{\text{ground}}|)}$$

where T_{pred} is the predicted fixation duration, and T_{ground} denotes the ground-truth fixation duration;

$$\text{FLA} = 1 - \frac{\sqrt{(X_{\text{pred}} - X_{\text{ground}})^2 + (Y_{\text{pred}} - Y_{\text{ground}})^2}}{\max\left(\sqrt{(X_{\text{pred}} - X_{\text{ground}})^2 + (Y_{\text{pred}} - Y_{\text{ground}})^2}\right)}$$

where X_{pred} and Y_{pred} respectively indicate the predicted fixation of x and the predicted fixation of y, and X_{ground} and Y_{ground} indicate the ground-truth fixation of x and the ground-truth fixation of y. Whenever the model perfectly predicts a scanpath, both FDA and FLA are set to 1, hence avoiding dividing by zero.

A.2 Results

We report the full results obtained by replicating Deng et al.'s experiments with both their Eyettention and our ScanEZ; in addition, we detail our ablation analyses, as well as our experiments on EML, a dataset that is not used in Deng et al. (2023): scanpath prediction performance on CELER L1 is in Table 3, broken down by the three settings of Text, Part and P.T.; Table 4 shows the cross-dataset results, providing evidence for ScanEZ's superior generalization abilities compared to Eyettention; Table 5 focuses on the contribution of the pre-training and fine-tuning steps of ScanEZ, each removed in the Doc, Part and P.T. settings; Table 6 does the same but on EML.

In Section 5, we noted that the advantage of fine-tuning ScanEZ on CELER L1 reflects in all metrics. On EML, however, FLA increases without fine-tuning. We thus conducted a small-scale analysis to study this difference between datasets, hypothesizing that scanpath length (i.e., the number of fixations) has an effect on FLA, as a space-related metrics. With a regression study, we found that this is indeed the case. Taking scanpath length as the independent variable and FLA as the dependent variable, we observed a significant effect across all datasets, including ZuCo 1.0: the positive t-scores in Table 7 denote that that longer scanpaths result

in higher FLA values. Notably, the t-score is higher on CELER L1 (92.36) than on EML (73.03), which indicates that longer scanpath sequences were easier to predict on the former dataset. This insight confirms that the more challenging data of EML make FLA improvements harder. Interestingly, the same conclusion can be drawn for FDA, although the increase of this metric in the "w/o Fine-tuning" setup was stable across datasets (cf. Table 1). Further analyses are therefore needed to better compare the two metrics, particularly regarding their relationship to scanpath features other than length.

Task	Model	NLL↓	NLD↓	NLL_t↓
Text	Eyettention	2.277 ± 0.005	0.572 ± 0.002	—
	ScanEZ	**1.603 ± 0.017**	**0.430 ± 0.047**	**1.269 ± 0.068**
Part	Eyettention	2.267 ± 0.005	0.573 ± 0.002	—
	ScanEZ	**1.555 ± 0.013**	**0.424 ± 0.025**	**1.235 ± 0.072**
P.T.	Eyettention	2.297 ± 0.011	0.568 ± 0.004	—
	ScanEZ	**1.524 ± 0.042**	**0.421 ± 0.022**	**1.244 ± 0.027**

Table 3: Performance on CELER L1 across the three split settings. Our model, ScanEZ, improves NLL and NLD and it benchmarks temporal predictions (see NLL_t scores, unavailable for Eyettention).

Training Data	Fine-tune	Testing Data	Model	NLL↓	NLD↓	NLL_t↓
ZuCo 1.0	—	ZuCo 1.0	Eyettention	2.653± 0.020	—	—
CELER L1	—			3.060± 0.026	—	—
CELER L1	ZuCo 1.0			2.613± 0.019	—	—
ZuCo 1.0	—	ZuCo 1.0	**ScanEZ**	**1.098± 0.072**	**0.608± 0.013**	**1.461± 0.035**
CELER L1	—			**0.829± 0.057**	**0.718± 0.008**	**1.214± 0.018**
CELER L1	ZuCo 1.0			**0.548± 0.069**	**0.690± 0.018**	**1.156± 0.030**

Table 4: Cross-dataset results following the training-testing set combinations of (Deng et al., 2023). Our model demonstrates better transfer performance than Eyettention based on NLL. We further include NLD and NLL_t scores, which are not reported in Eyettention.

Task	Model	NLL↓	NLD↓	NLL_t↓	FDA↑	FLA↑
Text	w/o Fine-tuning	3.021 ±0.044	0.952 ±0.004	1.406 ±0.027	0.604 ± 0.006	0.285 ± 0.016
	w/o Pre-training	1.715 ±0.254	0.537±0.065	1.419 ±0.259	0.634 ±0.021	0.376 ±0.050
	ScanEZ	**1.603 ±0.017**	**0.430 ±0.047**	**1.269 ±0.068**	**0.653 ±0.011**	**0.408 ±0.008**
Part	w/o Fine-tuning	3.032 ±0.223	0.951 ±0.015	1.409 ±0.073	0.604 ±0.011	0.286 ±0.024
	w/o Pre-training	1.791 ±0.204	0.517 ±0.104	1.536 ±0.292	0.623 ±0.021	0.378 ±0.045
	ScanEZ	**1.555 ±0.013**	**0.424 ±0.025**	**1.235 ±0.072**	**0.651 ±0.012**	**0.412 ±0.007**
P.T.	w/o Fine-tuning	3.035 ±0.204	0.950 ±0.009	1.403 ±0.012	0.603 ±0.004	0.284 ±0.018
	w/o Pre-training	1.772 ±0.304	0.547 ±0.085	1.431 ±0.150	0.611±0.054	0.381 ±0.054
	ScanEZ	**1.524 ±0.042**	**0.421 ±0.022**	**1.244 ±0.027**	**0.646 ±0.004**	**0.413 ±0.009**

Table 5: Ablation results on CELER L1. Removing pre-training or fine-tuning from ScanEZ degrades performance across all metrics. "w/o Pre-training" uses only human data; "w/o Fine-tuning" uses only E-Z Reader synthetic data.

Task	Model	NLL↓	NLD↓	NLL_t↓	FDA↑	FLA↑
Text	w/o Pre-train	1.331± 0.185	0.780± 0.003	1.164± 0.015	0.793± 0.008	0.502± 0.006
	w/o Fine-tune	1.760± 0.072	0.759± 0.004	1.191± 0.014	0.801± 0.005	**0.588± 0.004**
	ScanEZ	**1.000± 0.012**	**0.589± 0.012**	**1.106± 0.022**	**0.803± 0.004**	0.553± 0.020
Part	w/o Pre-training	1.154± 0.176	0.751± 0.062	1.151± 0.029	0.795± 0.007	0.520± 0.041
	w/o Fine-tuning	1.760± 0.070	0.759± 0.004	1.191± 0.018	0.801± 0.007	**0.588± 0.008**
	ScanEZ	**0.928± 0.037**	**0.585± 0.014**	**1.098± 0.026**	**0.803± 0.006**	0.556± 0.013
P.T.	w/o Pre-training	1.217± 0.212	0.777± 0.004	1.177± 0.066	0.785± 0.012	0.491± 0.023
	w/o Fine-tuning	1.837± 0.025	0.761± 0.005	1.181± 0.011	0.797± 0.003	**0.589± 0.006**
	ScanEZ	**0.996± 0.020**	**0.616± 0.013**	**1.083± 0.033**	**0.804± 0.005**	0.534± 0.012

Table 6: Evaluation on EML. "w/o Pre-training" uses only human data, "w/o Fine-tuning" uses only E-Z Reader synthetic data. ScanEZ is pre-trained on synthetic data and fine-tuned on human data.

Metric	CELER L1	ZuCo 1.0	EML
FDA	92.36	64.46	73.07
FLA	102.0	38.79	53.25

Table 7: The t-scores of the regressions performed for each dataset. All the scores are significant at p-value < 0.001.