Mouse-Guided Gaze: Semi-Supervised Learning of Intention-Aware Representations for Reading Detection

Seongsil Heo
Department of Computer Science and Engineering
University of California, Santa Cruz
Santa Cruz, CA 95064
sheo1@ucsc.edu

Roberto Manduchi
Department of Computer Science and Engineering
University of California, Santa Cruz
Santa Cruz, CA 95064
manduchi@soe.ucsc.edu

Abstract

Understanding user intent during magnified reading is critical for accessible interface design. Yet magnification collapses visual context and forces continual viewport dragging, producing fragmented, noisy gaze and obscuring reading intent. We present a semi-supervised framework that learns intention-aware gaze representations by leveraging mouse trajectories as weak supervision. The model is first pretrained to predict mouse velocity from unlabeled gaze, then fine-tuned to classify reading versus scanning. To address magnification-induced distortions, we jointly model raw gaze within the magnified viewport and a compensated view remapped to the original screen, which restores spatial continuity across lines and paragraphs. Across text and webpage datasets, our approach consistently outperforms supervised baselines, with semi-supervised pretraining yielding up to 7.5% F1 improvement in challenging settings. These findings highlight the value of behavior-driven pretraining for robust, gaze-only interaction, paving the way for adaptive, hands-free accessibility tools.

1 Introduction

Modeling human behavior from multimodal signals is a central goal in human-computer interaction. Among such signals, eye movements play a critical role in understanding how users engage with text [14, 21, 22]. In screen magnification environments, the visible area is restricted to only a few words or lines at a time [2]. To follow the text, users continuously drag the viewport with the mouse, moving horizontally to keep the current line in view and vertically to reach the next line. As a result, the mouse becomes indispensable yet taxing, especially for people with low vision who rely on screen magnifiers [16, 17]. This burden motivates automatic scroll control to reduce effort and preserve reading flow.

Intuitive and precise automatic scroll control requires accurate, low-latency inference of user intent. A key step is distinguishing focused reading from exploratory scanning. This distinction helps systems determine when and how to scroll based on the user's reading state, enabling more structured control

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundation Models for the Brain and Body.

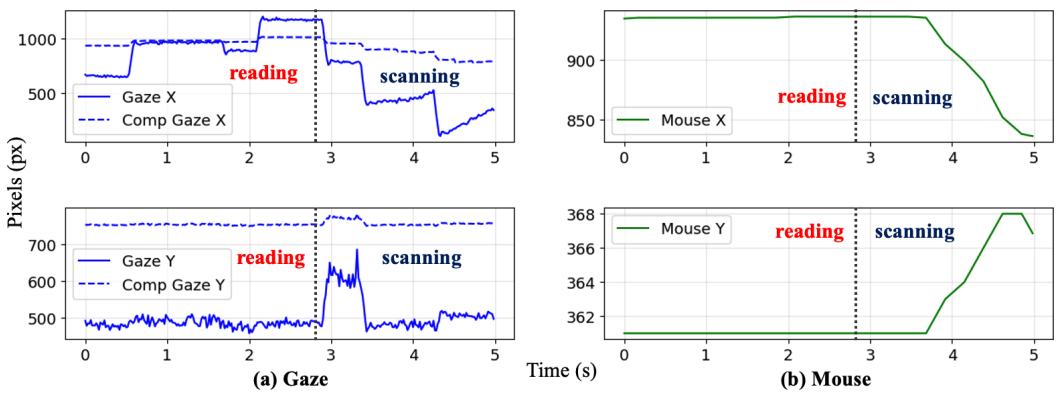

Figure 1: Five second segment from our dataset that spans a transition from reading to scanning (marked by a vertical dashed line). (a) Gaze position. Top panel shows X, bottom panel shows Y. Raw gaze is plotted with a solid line and the compensated gaze with a dashed line; "Comp" denotes compensated. (b) Mouse cursor position. Top panel shows X, bottom panel shows Y, with motion increases during scanning.

policies that adapt to behavioral context [25]. Although several gaze-based automatic scroll control systems have been proposed, most rely on handcrafted heuristics that generalize poorly across users and content types [18, 23]. To address these limitations, we propose a behavior classification model that robustly separates reading from scanning using gaze signals.

Distinguishing between reading and scanning from gaze alone is challenging, especially under screen magnification. In such settings, gaze trajectories are often fragmented and noisy due to limited visual context, viewport shifts, and disrupted eye movement patterns [6, 18, 20]. Fig. 1 (a) illustrates an example segment of gaze and compensated gaze. Gaze captures local eye movements relative to the magnified viewport, while compensated gaze remaps these coordinates back to the original screen space using the magnification factor[6, 19]. As shown, raw gaze appears irregular and noisy, while the compensated trace forms smoother, line-aligned trajectories, underscoring the difficulty of the task and the need for robust modeling.

To improve robustness, we use two complementary views of gaze. We jointly model raw and compensated gaze for behavior classification, capturing fine-grained movements while maintaining global spatial consistency in magnified settings. To our knowledge, this is the first approach to fuse these two streams for this task.

In addition, we introduce a semi-supervised learning framework that leverages mouse input only during training to guide the learning of gaze representations. As shown in Fig. 1 (b), mouse activity tends to increase during scanning episodes, reflecting users' need to reorient or reposition the viewport. Because mouse trajectories reflect deliberate, task-driven actions [9], they provide informative supervision even when gaze is noisy. Concretely, we pretrain the model to predict mouse velocity from unlabeled gaze and then fine-tune it for intent classification on a labeled set. At inference, it operates solely on gaze, enabling real-time, hands-free classification for accessibility.

2 Related Work

2.1 Gaze Remapping under Screen Magnification

Screen magnification restricts the visible area and induces systematic shifts in gaze coordinates. We adopt a prior remapping algorithm [6] to project gaze to unmagnified screen coordinates, yielding a consistent representation across magnification settings. This remapping recovers the global layout and preserves line and paragraph continuity that would otherwise be broken. We feed the compensated gaze to the model alongside raw gaze to capture complementary information: raw gaze preserves local oculomotor dynamics tied to the current viewport, whereas the compensated stream stabilizes trajectories in the screen reference frame.

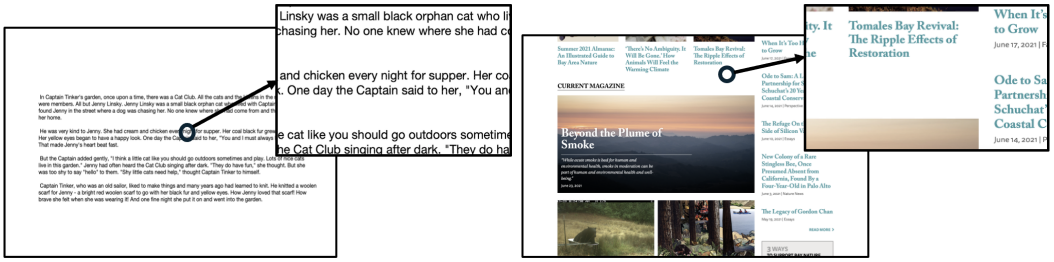

Figure 2: Examples of the two reading tasks (text document and webpage) under the full-lens magnification condition [24]. In this modality, the entire screen is uniformly enlarged, and only a portion of the content is visible at a time, requiring users to continuously scroll to follow the text.

2.2 Models for Reading Strategy Classification

Prior work on reading strategy classification has largely relied on hand-crafted features and pre-segmented text. For instance, baselines from ZuCo [8] required predefined sentence boundaries and computed aggregate gaze features over entire passages. While effective for post-hoc analysis, these approaches are not suited for consequent prediction or interaction. Our method avoids task-specific structure and learns temporal representations that enable fine-grained classification under naturalistic conditions. Statistical analyses of gaze metrics have also been used to compare reading strategies [5, 26], but these provide descriptive insights rather than trainable, generalizable models.

More recently, transformer-based models have been applied to classify reading-related behaviors such as reading, non-reading, and skimming from gaze and additional modalities, including head pose and RGB video [28]. These approaches demonstrate the promise of end-to-end sequence models for behavior classification, but they focus on normally sighted populations and do not consider the challenges introduced by screen magnification.

2.3 Self-supervised Learning for Gaze and Behavior

Recent advances in self-supervised learning have led to progress in gaze estimation from eye images [13, 27] and in coarse behavior recognition from signals like EOG [1, 11]. However, little work has been done on learning representations from continuous, high-resolution gaze trajectories. To our knowledge, no prior work has applied self- or semi-supervised learning to frame-level reading behavior classification from dense gaze sequences, particularly in accessibility contexts.

2.4 Mouse as Supervisory Signal during Reading

Mouse behavior has been shown to correlate strongly with gaze during reading, especially in screen-based environments. Cursor movements often reflect user attention and intent, making them a valuable implicit signal. Prior work has aligned mouse and gaze trajectories to analyze attention [12, 15] and incorporated cursor features into multimodal pipelines for predicting engagement and reading depth [3, 4, 10]. These studies highlight the potential of mouse input for reading behavior analysis, but they primarily focus on alignment or descriptive analysis.

3 Method

3.1 Dataset

We build on the dataset introduced by Tang et al.[24], which includes synchronized recordings of eye gaze and mouse trajectories from individuals with low vision during screen-based reading. In this work, we focus exclusively on the full-lens magnification condition, where the entire screen is magnified isotropically. Under this setting, users scroll right-to-left (often via left-to-right mouse movements), to keep the line being read within their preferred visual region. Participants read both text documents and webpages, as shown in Fig. 2 recorded with a Tobii Spectrum eye tracker (120Hz) and mouse logs (10Hz). Reading and scanning annotations were produced by trained annotators following a two-pass protocol under expert supervision.

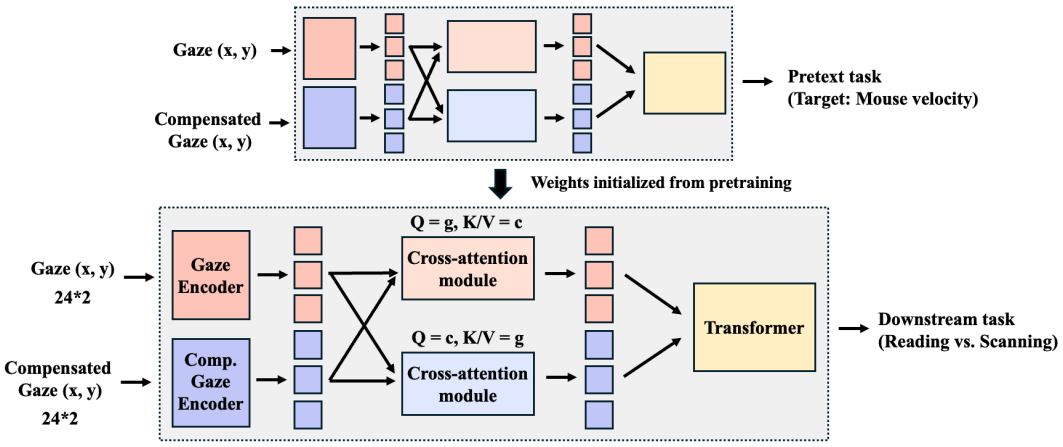

Figure 3: Overview of the framework. Top: pretraining (pretext). Bottom: fine-tuning & inference (downstream). "Comp." = compensated gaze; g = raw gaze, c = compensated. Both stages share the same backbone: g and c are fused via two cross-attention blocks (Q=g, K/V=c; Q=c, K/V=g) and processed by a Transformer. In pretraining, mouse movements serve only as targets to predict 2-D velocity from gaze; in downstream, the model classifies each sequence as reading or scanning.

3.2 Model

An overview of the supervised and semi-supervised frameworks is shown in Fig. 3. We adopted and extended the encoder and Transformer configuration from [28] and [7] to suit our task.

3.2.1 Supervised Classification

We first train a reading classifier on labeled gaze data. Input sequences are constructed using overlapping 0.2-second windows (24 steps at 120 Hz), where each window is labeled as reading or scanning based on its final time point. Each window contains two gaze streams: (1) raw gaze in magnified coordinates and (2) compensated gaze remapped to original screen space. Missing values are linearly interpolated, and windows with more than 50% missing samples are excluded.

Both gaze streams are independently encoded with a three-layer 1D CNN (kernel size 3, 64 dims) followed by cross-attention to capture complementary information. The fused representation is fed to a three-layer Transformer encoder(64 dims, 4 heads) to model temporal dependencies, and a linear classifier outputs the final label. Weighted cross-entropy is applied to address class imbalance.

3.2.2 Semi-Supervised Pretraining with Mouse Guidance

To enrich the gaze encoder with intention-aware features, we add a pretraining stage where the model learns to predict mouse velocity from gaze input. Mouse movements during magnified reading provide a weak supervisory signal that does not require manual behavior labels. The backbone encoder and Transformer are identical across pretraining and fine tuning. Only the task head and loss change: during pretraining, a linear regression head is trained with mean squared error to predict two-dimensional mouse velocity from unlabeled gaze, and during fine-tuning, the head is replaced with a classifier trained with cross entropy on intent labels.

4 Experiments and Results

To ensure subject-independent evaluation, we adopt a leave-one-subject-out cross-validation scheme, holding out each participant in turn for testing while training on the remaining subjects. All models were trained with the Adam optimizer (lr = 3e-4, weight decay = 0.01).

Table 1: F1 scores on the text dataset for different input configurations. "Comp." denotes compensated gaze. "Random" uses permuted labels with gaze-only input while preserving the original class distribution, serving as a sanity check. **Bold** indicates the best performance among gaze-only inputs.

Metric	*Gaze-Only Inputs*				*With Mouse Inputs*	
	Gaze + Comp.	Gaze	Comp.	Random	Mouse	Mouse + Gaze + Comp.
Overall F1	**80.02**	75.06	67.22	40.91	52.85	83.64
Reading F1	**91.27**	89.31	87.69	56.04	70.29	91.17
Scanning F1	**68.78**	60.81	46.75	25.79	35.41	76.10

Table 2: F1 scores of supervised and semi-supervised under two strategies (partial vs. full). 'Partial' updates only the last three Transformer layers, while "Full" fine-tunes the entire model. **Bold** indicates the best performance.

Input Type	**Text**			**Webpage**		
	Overall	Reading	Scanning	Overall	Reading	Scanning
Semi-supervised (Partial)	81.93	91.56	72.29	64.51	62.59	66.42
Semi-supervised (Full)	**85.97**	**93.13**	**78.80**	**70.01**	**68.39**	**71.62**
Supervised	80.02	91.27	68.78	62.49	60.27	64.59

4.1 Supervised Ablation on Gaze and Mouse Inputs

Supervised Learning with Gaze Input Table 1 compares three input configurations. Raw gaze and compensated gaze provide complementary views: raw captures fine-grained oculomotor dynamics within the magnified viewport, while compensated gaze restores global spatial consistency. Their fusion achieves the best gaze-only performance (80.02 F1), substantially higher than raw-only (75.06) or compensated-only (67.22). We select the eye (left or right) with the lowest NaN ratio in the first 10% of the session as a lightweight calibration heuristic. The random-label baseline (40.91 F1) confirms non-trivial of the task.

Supervised Learning with Mouse Input We first verify that mouse trajectories themselves provide meaningful behavioral signals. Mouse-only input is weaker than gaze (52.85 F1), yet combining it with gaze and compensated gaze improves performance to 83.64 F1. This shows mouse trajectories carry useful behavioral cues. Scanning F1 in particular improves markedly (68.78 $\rightarrow$ 76.10), indicating that mouse trajectories provide strong cues for exploratory behavior such as line skipping or reorientation. Since our target is hands-free interaction, however, mouse input is not available at inference. This motivates a semi-supervised approach in which mouse data serves only as auxiliary supervision during pretraining.

4.2 Semi-Supervised Gaze Representation Learning under Mouse Supervision

We pretrain with a mouse-guided gaze reconstruction objective and fine-tune on labeled data using two strategies: *partial*, which updates only the top three Transformer layers, and *full*, which updates all parameters. Both variants outperform the supervised baseline, demonstrating the utility of mouse signals as weak supervision. Even partial adaptation exceeds supervised training, underscoring the quality of the learned representations.

On text documents, full fine-tuning improves overall F1 by 6.0% (80.02 $\rightarrow$ 85.97). On the more challenging webpage setting, gains are larger: 7.5% (62.49 $\rightarrow$ 70.01). These results show that mouse-guided pretraining yields transferable gaze encoders, with benefits amplified under complex, variable environments.

5 Conclusion

We introduced a semi-supervised learning framework that uses mouse behavior to guide the learning of gaze representations for reading behavior classification under screen magnification. By pretraining on mouse-guided prediction and fine-tuning on labeled gaze, our method substantially improves over fully

supervised baselines. The joint modeling of raw and compensated gaze further enhances robustness, capturing both fine-grained oculomotor signals and consistent global patterns. Experiments on both text and webpage tasks demonstrate that mouse-guided pretraining is especially beneficial in noisy, complex environments. Importantly, inference remains gaze-only, making the approach practical for real-world, hands-free accessibility applications.

For future work, we plan to translate our intent classifier into a real-time, fully automatic scroll controller that interprets reading and scanning probabilities to generate smooth, safe viewport movements, leveraging hysteresis and uncertainty-aware mechanisms to enable robust hands-free accessibility on standard devices.

Acknowledgement

Special thanks to Tejas Polu, Aswhin Nagarajan, Suhas Oruganti, Shalini Raval, and Arav Adhikari for their assistance with data annotation and code debugging.

References

[1] Sriman Bidhan Baray, Mosabber Uddin Ahmed, Muhammad E. H. Chowdhury, and Koichi Kise. Eog-based reading detection in the wild using spectrograms and nested classification approach. *IEEE Access*, 11:105619–105632, 2023.

[2] Paul Blenkhorn, Gareth Evans, Alasdair King, S Hastuti Kurniawan, and Alistair Sutcliffe. Screen magnifiers: Evolution and evaluation. *IEEE Computer Graphics and Applications*, 23(5):54–61, 2003.

[3] C. Cepeda, A. Holzinger, P. Kieseberg, A.M̃. Tjoa, and E. Weippl. Mouse tracking measures and movement patterns with application for online surveys. In *CD-MAKE 2018: Computers Helping People with Special Needs*, volume 11015 of *Lecture Notes in Computer Science*, pages 28–42. Springer, 2018.

[4] Molly Conlen, Anand Kale, and Jeffrey Heer. Capture & analysis of active reading behaviors for interactive articles on the web. *Computer Graphics Forum*, 38(3):687–698, 2019.

[5] Benjamin Gagl, Klara Gregorova, Julius Golch, Stefan Hawelka, Jona Sassenhagen, Alessandro Tavano, David Poeppel, and Christian J Fiebach. Eye movements during text reading align with the rate of speech production. *Nature human behaviour*, 6(3):429–442, 2022.

[6] Seongsil Heo, Roberto Manduchi, and Suzana Chung. Reading with screen magnification: Eye movement analysis using compensated gaze tracks. In *Proceedings of the 2024 Symposium on Eye Tracking Research and Applications*, pages 1–6, 2024.

[7] Seongsil Heo, Calvin Murdock, Michael Proulx, and Christi Miller. Gaze-enhanced multimodal turn-taking prediction in triadic conversations. *arXiv preprint arXiv:2505.13688*, 2025.

[8] Nora Hollenstein and Caiming Zhang. Zuco 2.0: A multimodal dataset for reading and annotation tasks. *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4162–4169, 2020.

[9] Jeff Huang, Ryen W White, Georg Buscher, and Kuansan Wang. Improving searcher models using mouse cursor activity. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 195–204, 2012.

[10] Jialun Huang, Ryen W. White, and Georg Buscher. User see, user point: gaze and cursor alignment in web search. In *CHI '12: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1341–1350. ACM, 2012.

[11] Md. Rabiul Islam, Shuji Sakamoto, Yoshihiro Yamada, Andrew Vargo, Motoi Iwata, Masakazu Iwamura, and Koichi Kise. Self-supervised learning for reading activity classification. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(3):1–27, 2021.

[12] Aleksandar Jevremović, Panayiotis Zaphiris, Saša Adamović, Mati Mottus, and Andri Ioannou. Applying information theory and entropy to eliminate errors in mouse-tracking results. *arXiv preprint arXiv:2105.06320*, 2021.

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

[14] Johanna K Kaakinen, Ugo Ballenghein, Geoffrey Tissier, and Thierry Baccino. Fluctuation in cognitive engagement during reading: Evidence from concurrent recordings of postural and eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(10):1671, 2018.

[15] Ilan Kirsh and Mike Joy. Exploring pointer assisted reading (par): Using mouse movements to analyze web users' reading behaviors and patterns. In *HCI International 2020 – Late Breaking Papers: Multimodality and Intelligence*, volume 12424 of *Lecture Notes in Computer Science*, pages 156–173. Springer, 2020.

[16] Hae-Na Lee, Vikas Ashok, and IV Ramakrishnan. Bringing things closer: Enhancing low-vision interaction experience with office productivity applications. *Proceedings of the ACM on Human-computer Interaction*, 5(EICS):1–18, 2021.

[17] Seunghyun Lee and Jon Sanford. Gesture interface magnifiers for low-vision users. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*, pages 281–282, 2012.

[18] Roberto Manduchi and Susana Chung. Gaze-contingent screen magnification control: A preliminary study. In *International Conference on Computers Helping People with Special Needs*, pages 380–387. Springer, 2022.

[19] Roberto Manduchi, Seongsil Heo, and Susana TL Chung. Eye movement analysis for low vision readers using a full screen magnifier. *Investigative Ophthalmology & Visual Science*, 65(7):1117–1117, 2024.

[20] Natalie Maus, Dalton Rutledge, Sedeeq Al-Khazraji, Reynold Bailey, Cecilia Ovesdotter Alm, and Kristen Shinohara. Gaze-guided magnification for individuals with vision impairments. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2020.

[21] Brian W Miller. Using reading times and eye-movements to measure cognitive engagement. *Educational psychologist*, 50(1):31–42, 2015.

[22] Gary E Raney, Spencer J Campbell, and Joanna C Bovee. Using eye movements to evaluate the cognitive processes involved in text comprehension. *Journal of visualized experiments: JoVE*, (83):50780, 2014.

[23] Selina Sharmin, Oleg Špakov, and Kari-Jouko Räihä. Reading on-screen text with gaze-based auto-scrolling. In *Proceedings of the 2013 Conference on Eye Tracking South Africa*, pages 24–31, 2013.

[24] Meini Tang, Roberto Manduchi, Susana Chung, and Raquel Prado. Screen magnification for readers with low vision: A study on usability and performance. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–15, 2023.

[25] Jayson Turner, Shamsi Iqbal, and Susan Dumais. Understanding gaze and scrolling strategies in text consumption tasks. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, pages 829–838, 2015.

[26] Sebastian Wallot. *The role of reading fluency, text difficulty and prior knowledge in complex reading tasks.* University of Cincinnati, 2011.

[27] Yong Wu, Gongyang Li, Zhi Liu, Mengke Huang, and Yang Wang. Gaze estimation via modulation-based adaptive network with auxiliary self-learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):5510–5520, 2022.

[28] Charig Yang, Samiul Alam, Shakhrul Iman Siam, Michael J Proulx, Lambert Mathias, Kiran Somasundaram, Luis Pesqueira, James Fort, Sheroze Sheriffdeen, Omkar Parkhi, et al. Reading recognition in the wild. *arXiv preprint arXiv:2505.24848*, 2025.