Reading Recognition in the Wild

Charig Yang[1,2], Samiul Alam[3], Shakhrul Iman Siam[3], Michael J. Proulx[1], Lambert Mathias[1], Kiran Somasundaram[1], Luis Pesqueira[1], James Fort[1], Sheroze Sheriffdeen[1], Omkar Parkhi[1], Carl Ren[1], Mi Zhang[3], Yuning Chai[1], Richard Newcombe[1], Hyo Jin Kim[1]

[1]Meta Reality Labs Research [2]VGG, University of Oxford [3]The Ohio State University
charig@robots.ox.ac.uk, kimhyojin@meta.com

https://www.projectaria.com/datasets/reading-in-the-wild/

Abstract

To enable egocentric contextual AI in always-on smart glasses, it is crucial to be able to keep a record of the user's interactions with the world, including during reading. In this paper, we introduce a new task of *reading recognition* to determine *when* the user is reading. We first introduce the first-of-its-kind large-scale multimodal *Reading in the Wild* dataset, containing 100 hours of reading and non-reading videos in diverse and realistic scenarios. We then identify three modalities (egocentric RGB, eye gaze, head pose) that can be used to solve the task, and present a flexible transformer model that performs the task using these modalities, either individually or combined. We show that these modalities are relevant and complementary to the task, and investigate how to efficiently and effectively encode each modality. Additionally, we show the usefulness of this dataset towards classifying types of reading, extending current reading understanding studies conducted in constrained settings to larger scale, diversity and realism. Code, model, and data will be public.

Figure 1: **Am I reading?** The left figure shows a timeline as the user navigates the world. We aim to solve the task of reading recognition to enable AI assistants in always-on wearables. We identify three modalities: eye gaze (in colored dot patterns), RGB crop around gaze (in red box), and inertial sensors performs the task to high accuracy (with Prediction and GT shown). Images from our *Reading in the Wild* dataset, which features 100 hours of diverse reading and non-reading activities in real-world settings, with examples shown in the right.

1 Introduction

The potential future of AI personal assistants depends on its ability to understand the physical context of the user. Smart glasses are becoming a promising device form factor capable of linking visual AI capabilities to the real world. Recently, there has been a sharp rise in the development of smart glasses, both products (Meta Ray-Ban, Amazon Echo Frames) and prototypes (Snapchat Spectacles, Halliday AI Glasses, Xreal One Pro). These all-day wearable devices enable proactive, personalized, and contextualized AI agents to perceive the world like humans do by understanding the users' context.

However, for always-on wearable glasses, due to both *hardware* (power, bandwidth, heat) and *software* (perception capability of AI agents, especially with heavy models) constraints, it is impractical to

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

record and process every single frame over long periods of time. One solution is to have a proxy signal, so that the device can record and process *key* frames only *when* relevant. The question becomes: what forms important context of the user that the AI assistant needs to know, and how do we know when to capture them?

The ability to read underpins one of, if not the most important unique modalities by which modern humans communicate, entertain each other, and learn. Reading is a key mechanism humans use to communicate with high fidelity and high information density. Reading spans a broad array of mediums, from handwritten and printed text on paper and digital displays to environmental signposts. The act of reading occurs within real-time communication with one another and today's AI chatbots, through to reading long-form articles in books or online. Enabling AI with the ability to recognize reading is hence clearly one of the most important context signals a future AI can be enabled with to unlock truly personalized and contextually relevant AI.

Given this, we ask: how can we provide future AI with the ability to know when someone is reading? This apparently simple idea underpins the ability to efficiently enable devices to know what the user has and has not read, and hence where they can assist given what it understands that the user has read.

This task of *reading recognition* is challenging for two main reasons. First, the problem can often be ill-posed: just because a text exists in the field of view does not mean that the user is reading it (or even looking at it), which is ambiguous to solve using visual information alone. Also, the method should be efficient for real-time, always-on computation subject to the practical constraints of a wearable device. Both of these challenges render OCR-based text detection methods impractical, given the inability to solve the ambiguity and the requirements for high-resolution capture and processing. Instead, reading recognition can be used as an efficient proxy to indicate *when* and *where* it is relevant to invoke heavier models (OCR and VLMs) instead of running these models all the time.

Motivated by this question, we introduce a new dataset created with Project Aria [11] glasses, which enables us to develop the contextual AI capability of detecting *when* a wearer is reading. We present the first-of-its-kind large-scale multimodal "Reading in the Wild" dataset, containing 100 hours of reading and non-reading videos in diverse and realistic scenarios. This dataset allows us to identify three modalities (egocentric RGB, eye gaze, head pose) that can be used to solve the task. We then present a flexible transformer model that performs the task using these modalities, either individually or combined. We show that these modalities are relevant and complementary to the task and investigate how to efficiently and effectively encode each modality, as well as the model's ability to generalize towards unseen scenarios and perform real-time reading detection.

Achieving reading recognition makes it feasible to keep a record of a user's reading interactions with the world to build a contextually aware AI. It also enables several other applications: it allows reading assistant tools [38] in children with learning difficulties [5] and people with low vision [40] to operate in the real world; it can also be used to track whether a user has read crucial information (*e.g.* signs during driving) and to measure attention and distraction while performing a task.

Additionally, the dataset and method contributed in this paper can be extended to classifying different types of reading. This has been of interest in cognitive studies in reading comprehension, but they are often limited to controlled environments [25, 26, 21, 1, 7], hence limiting its usefulness. We show that our dataset allows for reading mode and medium classification to be performed in unconstrained settings, and provide experimental results in this direction.

In summary, we make the following contributions:

- First, we introduce a new task of reading recognition *in the wild*, and demonstrate its usefulness. Unlike previous studies, we focus on in-the-wild settings and practicality towards wearable glasses.
- Second, we present the first-of-its-kind large-scale egocentric multimodal *Reading in the Wild* dataset, which will be made publicly available, alongside a scalable protocol for data collection.
- Third, we identify three modalities relevant and complementary to the task (RGB, gaze, and IMU), and develop a lightweight, flexible model that inputs these modalities either individually or in combination for reading recognition, resulting in a strong and efficient baseline for this task.
- Fourth, we show that our method and dataset extend towards reading understanding, including classifying reading mode and medium, demonstrating usefulness towards cognitive studies.

Subset	Size	Indoor	Outdoor	Medium	Text type	Multi-task	Mode	Language	Not reading	Mixed
Seattle (train/val/test) **Focus**: diversity	80 hours 81 people 1061 videos	Offices Libraries Homes Stores	Balconies Patios Roads/trails In the woods	Print Digital Objects	Paragraphs Short texts Non-texts Dynamic texts	None Walking Writing Typing	Engaged Skimming Scanning Out loud	English (→)	Daily activities Hard negatives (71%/29%)	Alternating sequences (reading / not reading)
Columbus (test) **Focus**: edge cases, generalization	20 hours 31 people 655 videos	Offices Libraries Lounges Corridors		Print Digital Objects	Paragraphs Short texts Non-texts	None	Engaged Scanning	English (→) Bengali (→) Chinese (↓) Arabic (←)	Hard negatives Daily activities (58%/42%)	Mirror setups (same settings, one reading, another not)

Table 1: **Dataset overview.** We separately collect two subsets for the dataset. Seattle subset focuses on diversity, while Columbus subset looks at the model's generalization towards unseen settings, as well as edge cases where the model fails. See Appendix A for more details.

2 Related Work

Reading recognition has been a long studied task with rich literature. Eye gaze has been used as the primary signal [21, 6, 1, 26], however, it relied on handcrafted feature engineering methods such as detecting fixations and saccades, which we show are unnecessary. Moreover, the experiments are usually constrained, and not performed *in the wild*. Other modalities have also been considered, such as electrooculography (EOG) signals [4], though the usage of electrodes can be invasive and hence less practical towards building user-friendly wearable glasses. In this paper, we steer this towards practical usage in modern smart glasses, where we show that gaze can be used in combination with visual information and IMU sensors. With recent advances in wearable devices, reading recognition expands to tasks such as word recognition and reading order prediction [17]. While this is relevant, it concerns the reading content, and assumes the user is already reading, which differs from the task of detecting whether the user is reading in this paper. Applications include reading comprehension [31, 19, 9, 36], understanding user behavior [7], and in building reading assistants [38, 5, 40]. However, the literature is largely constrained to controlled environments.

Egocentric activity recognition is a popular vision task that usually require computationally heavy solutions using video input [20, 45]. In terms of data, reading is only a subset of activities in some common datasets such as EGTEA Gaze+ [27] and Ego-Exo4D [14]. However, not all datasets contain reading [10]. For those which include reading, its nature is very restricted to activities such as reading recipes (in [27]), reading covid testing manuals, climbing instructions, and music sheets (in [14]). Ego4D [13] offers a more diverse range of reading activities, but only less than 1% of the data includes eye gaze. In contrast, our paper focuses on efficient reading recognition, and the proposed dataset contains large-scale and diverse reading and non-reading examples with eye gaze information.

Gaze in computer vision has started to gain popularity, where gaze has many applicable uses. One popular route is to perform gaze prediction *i.e.* predicting where the user is looking at [34, 12, 32, 37] or how the user interacts with objects he/she observes [18, 39, 29]. In medical applications, eye gaze can be used as a saliency test to ensure integrity in medial image analyses [41, 23, 22, 30], as well as predicting learning disorders [16]. Recently, gaze has also been used to complement vision, such as in action recognition [44], narration [8], and vision-language models [24]. Our paper further explores whether gaze can *reduce* the input requirements for computer vision models by only using gaze and/or parts of vision that are associated with gaze instead of using the whole image sequence.

3 Reading in the Wild Dataset

3.1 Overview

The dataset contains about 100 hours of recordings of reading and non-reading activities collected from one RGB (30Hz, 1408p, 110° FoV) and two SLAM (150° FoV) cameras, two eye tracking cameras (60Hz, calibrated), two IMUs (with odometry outputs from visual SLAM), and audio transcribed using WhisperX [2]. We independently collect two subsets of this dataset, as in Table 1.

Seattle is collected for training, validation, and testing. We mainly focus on collecting reading and non-reading activities in diverse scenarios, in terms of participants' identities, reading scenarios, reading modes, and reading materials. It contains a mix of normal and hard examples, as well as mixed sequences alternating between reading and non-reading activities. The dataset is collected in homes, office spaces, libraries, and outdoors.

Columbus is collected to find out where the model breaks in zero-shot experiments. It contains examples of hard negatives (where text is present but is not being read), searching/browsing (which gives confusing gaze patterns), and reading non-English texts (where reading direction differs).

Dataset	Gaze	RGB	Reading	Real	HN
Ego4D	✗	✓	Limited	✓	✗
Ego-Exo4D	10Hz	✓	Limited	✓	✗
EGTEA	30Hz	✓	Limited	✓	✗
ZuCo	500Hz	✗	✓	✗	✗
InteRead	1.2kHz	✗	✓	✗	✗
Ours	60Hz	✓	✓	✓	✓

Table 2: **Comparison to existing datasets**. Our dataset is the first reading dataset that contains high-frequency eye-gaze, diverse and realistic egocentric videos, and hard negative (HN) samples.

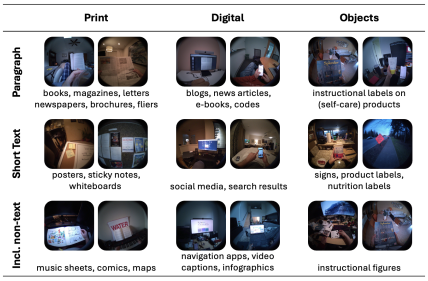


Figure 2: **Diversity in reading materials.** Reading examples across different materials, both text type (rows) and medium (column).

3.2 Comparison to existing datasets

The closest kins to our dataset come in two categories, as shown in Table 2. First, in egocentric video datasets [14, 13, 27], there are very limited reading sequences and they lack diversity as each dataset only reads from 1-2 examples (COVID test kits for Ego-Exo4D, recipes for EGTEA). Moreover, their eye tracking frequencies are also limited. Second, there are cognitive studies that focus on human gaze behavior during reading [15, 43] with high-frequency eye tracking. However, these studies are conducted in very constrained scenarios such as reading a text in front of a screen. Moreover, these studies only collect gaze data without RGB stream.

3.3 Contents

Reading. Our dataset presents a large diversity in reading activities, including:

- **Reading mode**: Our dataset contains different reading modes, including deep reading (careful, engaged reading), skimming (quickly glancing through for general ideas), scanning (searching for specific information), and reading aloud (verbalizing the text).
- **Single/Multi-task reading**: Our dataset not only covers single-task reading, where the focus is solely on the reading material, but also reading while multitasking, such as reading while writing, typing, or walking.
- **Medium and text type**: We collect data across mediums: print (books, newspapers, flyers), digital (phones, monitors), everyday objects (product labels, whiteboards); and text types: paragraphs, short texts, non-texts, and dynamic texts (video captions and subtitles) as illustrated in Figure 2.
- **Demographics**: We collect data among 111 participants and include their age range and gender.
- **Location**: For diversity, we collect scenes across indoor (e.g., meeting rooms, bedrooms, living rooms), balconies, outdoors, and in the woods.

Non-reading. We also collect negative examples. This includes *Everyday activities* that do not involve reading such as physical exercise, outdoor activities, creative arts, culinary activities, and household chores, as well as *Hard negatives*, where text is present in the scene but is not being read, which would confuse RGB-only models.

Mixed. We also collect *Alternating sequences*, where the participants alternating between reading and non-reading with annotated timestamps, and *Mirror setups* where we have the same participant perform reading and non-reading activity in the same environment and the same material.

3.4 Data collection process

Logistics. We recruited a total of 111 participants, targeting a uniform distribution for gender and age. We gave each participant a list of tasks to record, with moderators monitoring to ensure that the recordings are correct as desired.

Instructions. We divided the collections into tasks, each with specific instructions, as elaborated in the Appendix. We also asked the participants to perform eye gaze calibration within each recording.

Privacy. We strictly followed Project Aria Research guidelines. All data has been de-identified, and faces and license plates were anonymized with EgoBlur [35]. We source the venues ourselves do not use the participants' private spaces to prevent exposure of sensitive or identifiable information.

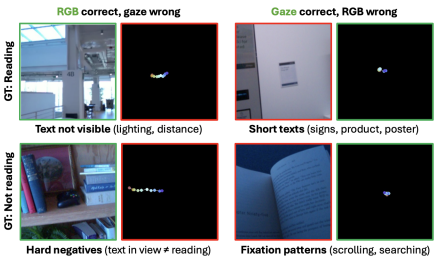


Figure 3: **Complementary modalities**. Example success and failure cases for gaze and RGB, suggesting the benefit of multimodality.

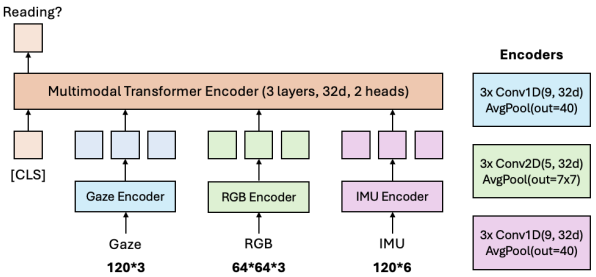


Figure 4: **Model architecture**. Our model is a simple transformer encoder with any combination of gaze, RGB, and IMU as input.

Scalable Protocol through Automatic labeling. In addition to the dataset itself, we also present a protocol for scalable, high-quality data collection. Instead of manually labeling the timestamps, we instruct the participants to say "start reading!" whenever they start reading, and "finished reading!" whenever they finish. In doing so, we can simply use WhisperX [2] to obtain accurate timestamps without requiring manual annotations.

Quality assurance. We have several protocols to ensure that participants have read the text. This involves before (pre-reading questions) and after (post-reading questions and summarization). For the subset where the user reads out loud, the audio transcription can also be used to for quality assurance.

4 Method

4.1 Task definition

Formally, at time t, we want to predict the confidence score $s_t \in [0, 1]$ whether the user is reading or not, given several input modalities: eye gaze patterns $g_{t-T\leq\tau\leq t} \in \mathbb{R}^{f\times T\times d}$, instantaneous RGB $I_t \in \mathbb{R}^{H\times W\times C}$ and head pose (IMU) sensor readings $z_{t-T\leq\tau\leq t} \in \mathbb{R}^{f\times T\times d}$, where f is the sampling frequency and T is the input duration *i.e.* $s_t = \Phi(g_{t-T\leq\tau\leq t}, I_t, z_{t-T\leq\tau\leq t})$. Each modality has different advantages and drawbacks. To harness the strength of all modalities, we propose a multimodal model that takes into account all three modalities as input. In the following sections, we first discuss individual modalities, followed by the model architecture.

4.2 Input modalities

Gaze. There exists a vast literature suggesting that gaze can be used to detect reading activity without visual information [21, 6, 1, 26]. However, their experiments are limited to constrained environments (reading long paragraphs in front of a screen), and they rely on feature engineering methods such as fixation detection to circumvent small-scale data. As we demonstrate in the experiments section, training on diverse data translates well to open-world settings, and feature engineering is unnecessary at scale, which makes it robust to low frequency eye tracking inputs.

RGB. As with action recognition methods, visual information has been an effective cue in the computer vision community. However, processing video models on a wearable device is expensive. Meanwhile, there has been an interest in using gaze to guide model attention in action recognition [27, 14, 44]. For reading, we argue that region outside the gaze point is likely to be irrelevant, as the high-resolution human fovea capable of reading only covers a small region (2°) around the gaze [33]. Therefore, we only crop the image around the gaze region. This also allows for large efficiency gains as capture and processing only needs to be done on a small patch. We find that cropping using only 1/484 of an image (64px, 5° from 110° FoV) can result in good accuracy, with the remainder for context and gaze uncertainties.

Head pose (IMU). We also explore using odometry measurements. While not a good indicator on its own, we find that it helps as a secondary sensor. The intuition here is that some inertial motions can be used to address ambiguities, such as distinguishing between reading and horizontal head motion.

Complementary modalities. The main reason for using multiple modalities is that they are complementary: they excel and fail in different places. For example, eye gaze can perform well even if the text is not visible due to lighting or distance that images sometimes miss out, while RGB works in cases where gaze patterns are not obvious, such as when reading short texts like signs, as

shown in Figure 3. While IMU is not strong on its own, we show later that it further provides cues to disambiguate some cases (e.g. turning heads vs reading).

4.3 Model

In order for this to be practical towards always-on wearable devices, we propose a simple and efficient model that achieves a strong practical baseline for the task. Particularly, we propose a flexible multimodal transformer model that takes in different modalities as input, as shown in Figure 4. By keeping the model simple, we can investigate different combinations and forms of modalities.

Input. Unless otherwise stated (such as in ablation studies), we use $T = 2$, $f = 60$ for 3D eye gaze and 6DoF IMU, and a 5° FoV ($H, W = 64$) crop for RGB as default.

Modality encoder. The model consists of different encoders $\Phi_{\{g,r,i\}}$ (where g,r,i represent gaze, RGB, and IMU respectively) to tokenize individual modality into feature tokens $f_{\{g,r,i\}} \in \mathbb{R}^{N \times D}$. We use three layers each of 1D (gaze and IMU) and 2D (RGB) convolutions.

Multimodal transformer. We then combine these feature tokens using a simple transformer encoder Φ_t and a linear head over the [CLS] token *i.e.* $s_t = \Phi_t(f_g, f_r, f_i)$.

Modality dropout. During training, we dropout entire modalities at random, which serves two purposes: (i) it helps with training less-used modalities; (ii) during inference, the model can perform well even without all modalities being present.

4.4 Generalization

While we train on English texts, we find that our model generalizes well to other left-to-right languages across different writing systems, but struggles with vertical and right-to-left texts, as the gaze pattern is in a different direction. To address this, we find that simply augmenting the gaze at inference time (90° rotation for vertical texts and horizontal flip for right-to-left texts) allows the model to generalize well. In practical scenarios, this can be done depending on geo-location. During training, we also add a small fraction of rotated gaze to help with reading vertical texts.

5 Experimental Setup

5.1 Dataset split

We split the Seattle subset into training, validation, and test sets, and train the model on the training set. We evaluate on (i) the test set of the Seattle subset, and (ii) the entire Columbus subset. We also evaluate on specific subsets to study latency and generalization.

5.2 Implementation details

Model. For the encoders, we use three layers of 1D convolution (kernel size 9, 32 dims) for gaze and IMU, and three layers of 2D convolution (kernel size 5, 32 dims) for RGB. We then feed the tokens as input to three layers of transformer encoder (32 dims, 2 heads) before linearly projecting the [CLS] token to two classes. The combined model is lightweight, with 137k parameters.

Training. We impose modality dropout such that there is an equal probability of using one, two, or three modalities at the same time, as well as perform rotation augmentation. We use Adam optimizer with learning rate $1e^{-3}$ for ten epochs. All models are trained using a single GPU. The code and models will be released alongside the dataset.

5.3 Evaluation metrics

Classification metrics. We calculate the accuracy and F1 scores for each task at 0.5 confidence threshold. We also vary this threshold, and report the precision at 0.9 recall (denoted as $P_{R=.9}$).

Latency. We consider latency to be the time between a state change and model detecting it, and is unrelated to the computational time, which we assume to be negligible given the small model size.

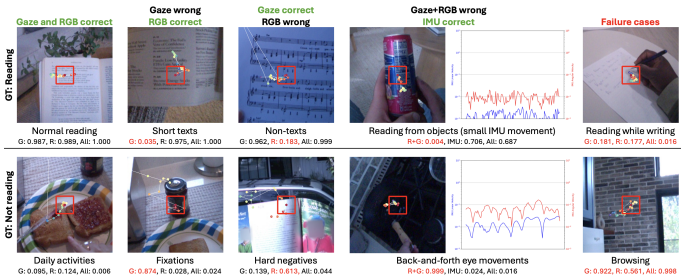
6 Results

6.1 Main results

We present the main results and visualizations in Figure 5.

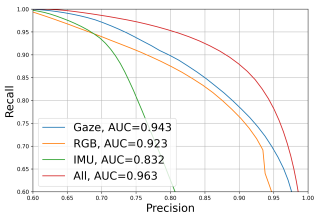
Gaze	RGB	IMU	Acc	F1	$P_{R=.9}$
✓			82.3	84.5	79.8
	✓		82.2	83.7	76.5
		✓	74.7	80.0	71.9
✓		✓	84.9	86.5	83.6
	✓	✓	83.5	85.2	82.3
✓	✓		86.0	87.8	87.3
✓	✓	✓	**86.9**	**88.1**	**88.0**

(a) Main results



(b) Visualization (**G**/**R** are for gaze/RGB, with wrong ones in red)

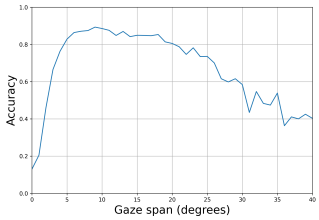
Figure 5: **Main results and visualizations.** We show the results on Seattle (test set). (a) Our method performs the task to good accuracy, and combining all modalities yields the best results. Metrics are accuracy and F1 score at 0.5 threshold, and precision at 0.9 recall. (b) We show: (i) Col. 1, banal success cases distinguishing reading from daily activities; (ii) Col. 2-4, difficult cases where our combined model predicts correctly even if individual modality fails, including reading from objects, short texts, non-texts, fixation patterns, and hard negatives; (iii) Col. 5, failure cases where all modalities fail, including reading while writing and browsing.



(a) Precision-recall curve

Scenario	Acc (%)
Digital media (normal scenarios)	95.3
Print media (normal scenarios)	93.8
Reading average	**88.1**
Objects (normal scenarios)	87.6
Reading while walking	81.4
Reading from videos	78.0
Reading non-texts	65.8
Reading while writing/typing	55.5
Daily activities	95.2
Not reading average	**86.4**
Hard negatives	74.7

(b) Breakdown by scenario



(c) Breakdown by gaze span

Figure 6: **Results breakdown.** We present the breakdown for the main results, including (a) precision-recall curve for different modalities (b) breakdown by scenario to highlight difficult cases (c) breakdown by gaze span.

Single modality. We find that gaze and RGB are able to achieve reasonable performance individually, and their performances are similar to each other (82.3% and 82.2% accuracy respectively). However, as shown in the visualizations, they have different success and failure cases. IMU alone does not perform very well, which is reasonable, as the problem becomes ill-posed, and the model can only guess the lack of motion as not reading (and vice versa).

Combined modalities. We find that IMU monotonically improves upon gaze (+2.6%) or RGB (+1.3%) as secondary modality, with small extra compute. Qualitatively, we see that IMU helps improve several corner cases, and RGB is particularly strong for short texts. We also find that all modalities combined yields the best performance of 86.9% in accuracy (+4.6% from best single-modality model), validating the complementary roles of different modalities.

6.2 Results breakdown

We show the breakdown of results in Figure 6.

Scenario breakdown. We break down the results of the combined model. We find that the model mostly succeeds in normal cases, but fails in cases where reading is atypical, such as reading non-texts (maps, music sheets), or when reading while writing or typing. The model also struggles with hard negative examples introduced in this dataset.

Gaze span breakdown. We also break down the results of reading sequences by the horizontal gaze field of view, as it correlates with text size. We find that the accuracy is the highest (86.1%) for fields of view of 5-20°, corresponding to 64-256 pixels, with accuracy dropping sharply for both below (59.3%) and above (70.6%) this range.

6.3 Generalization

We use the model trained on the Seattle subset to evaluate on unseen scenarios, shown in Table 3.

Zero-shot generalization. To evaluate zero-shot capabilities, we test on the separately collected Columbus subset. We show that the model performs reasonably zero-shot, and draw similar conclu-

Gaze	RGB	IMU	Acc	F1	$P_{R=.9}$
✓			77.1	84.0	84.1
	✓		76.7	84.5	83.4
✓	✓		82.8	88.7	**88.2**
✓	✓	✓	**82.9**	**88.8**	**88.2**

(a) Zero-shot on Columbus

Language	Aug	Acc	F1
English →	-	81.2	87.0
Bengali →	-	93.0	95.9
Chinese ↓	-	35.5	51.6
	rotate	85.1 (+49.6)	91.9 (+40.3)
Arabic ←	-	21.0	23.8
	flip	51.5 (+30.5)	63.8 (+40.0)

(b) Cross-language (text direction)

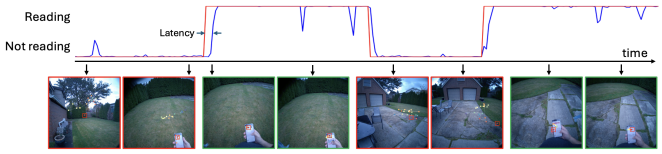
Test	Train	Acc	F1
Seattle	Seattle	79.3	81.2
	EGTEA	62.9 (-16.4)	56.9 (-24.3)
EGTEA	EGTEA	89.6	70.6
	Seattle	87.7 (-1.9)	63.4 (-7.2)

(c) Generalization to EGTEA

Table 3: **Generalization results.** Using model trained on Seattle subset, we test on (a) separately collected Columbus subset; (b) different languages with different reading patterns and direction (despite only being trained with English), where we explore using rotation and flipping augmentations; (c) cross-generalization with EGTEA. The model generalizes one way (Seattle → EGTEA) but not the other.

Gaze	RGB	IMU	Acc	F1	Latency (s)
✓(1s)			77.1	75.6	0.526
✓(2s)			79.0	78.9	0.831
✓(3s)			79.3	77.8	1.013
	✓		73.8	68.7	**0.321**
✓(2s)	✓		81.7	79.5	0.642
✓(2s)	✓	✓	**82.7**	**81.0**	0.720

(a) Latency



(b) Visualization using Gaze+RGB+IMU model. [Prediction/GT]

Figure 7: **Real-time detection.** We evaluate our model on alternating sequences for real-time detection. In (a), we show that (i) longer gaze sequences result in higher latency, (ii) RGB has lower latency than temporal signals (iii) adding RGB to gaze reduces the latency compared to gaze alone. We illustrate the results in (b).

sions in terms of the complementary role between gaze and RGB, but IMU does not help as much given that the dataset does not contain freeform daily activities where IMU helps the most.

Further, we also notice the differences in reading speed across different users, especially across different languages. It is possible to personalize the model by scaling the gaze to the magnitude of the reader, and empirically this solves some of the failure cases.

Cross-language generalization. While we only train the model on English, we find that our model generalizes well towards non-English, left-to-right texts, but less well on other languages where the reading direction is different. To circumvent this during inference, we perform 90° rotation to tackle vertical texts, and horizontally flip the gaze for right-to-left texts. We show that using gaze-only model solves the problem to a reasonable extent.

Cross-dataset generalization. To demonstrate the importance of collecting reading examples in freeform settings, we conduct experiments to test for generalizability across datasets. For this, we utilize EGTEA Gaze+ dataset [27], where we only use their 'reading' action labels, and treat other labels as not reading. To match the data available in EGTEA, we use 2D gaze projection at 30Hz. We conduct cross-generalization experiments where we train on one training set and evaluate on the other test set. We show that training on EGTEA with limited training samples does not generalize to in-the-wild scenarios, whereas the generalization gap for our dataset is much smaller.

6.4 Application: real-time reading detection

So far, we only consider atomic predictions to answer *whether* someone is reading. To extend to *when*, we simply perform predictions over time. To evaluate this task, we use the alternating sequences between reading and not reading with labeled timestamps, as shown in Figure 7. On top of the evaluation metrics, we also evaluate the latency (i.e. the duration required for a state change to be detected). Our results show that (i) there is a trade-off between gaze duration and latency; (ii) RGB has lower latency as the predictions are instantaneous, and does not rely on past detections; and (iii) combining gaze and RGB reduces the latency compared to gaze-only model.

Localization. To extend to *where* the user is reading, we can use the gaze point to locate the texts. As such, OCR only needs to be performed around the gaze, which results in additional compute savings. Also, the gaze scanpath can be used to estimate how much to crop the image for OCR.

Efficient interface for OCR. OCR comes in two phases: text detection and recognition. Using reading recognition as a low-compute interface allows OCR to run not as often, and on a smaller image each time. Furthermore, the reading detection model is designed to be small enough for on-device compute, so that images need to be transferred off-device only when reading detected, significantly reducing bandwidth requirements.

Input	Acc	F1	$P_{R=.9}$
Retina images	79.2	83.0	76.2
3D ray (d/dt)	82.1	84.2	78.4
3D point	80.8	83.3	77.9
3D point (d/dt)	**82.3**	**84.5**	**79.8**
2D projection	79.8	81.3	74.6
Gaze + IMU	83.9	85.7	80.0
Gaze + VIO	**84.9**	**86.5**	**83.6**

(a) Input representation

Freq	Acc	F1	Dur	Acc	F1
60	**82.3**	**84.5**	5	**85.8**	**87.5**
30	81.7	84.3	4	85.4	87.1
20	81.3	83.6	3	83.6	85.7
10	80.4	82.9	2	82.3	84.5
6	79.2	82.0	1	79.6	82.2

(b) Gaze frequency and duration

FoV	Acc	F1
14	**83.5**	**85.1**
10	82.9	84.6
7	82.9	84.3
5	82.2	83.7
3.5	79.5	80.6

(c) RGB crop size

Model	Acc	F1
XS (6k)	82.0	83.6
S (34k)	86.3	87.7
M (137k)	86.9	88.1
L (600k)	87.1	88.8
XL (1M)	**88.5**	**90.1**

(d) Model size

Table 4: **Ablation studies.** We show ablation studies for (a) the representations for gaze and IMU, (b) the gaze frequency and duration, (c) RGB crop size, and (d) model size. We fix other experiments to 60Hz, 2s, and 5° FoV using the M (137k) model, as underlined.

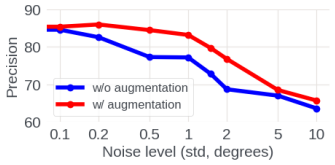


Figure 8: **Noise robustness.** Augmentation (red) lowers degradation.

GT \Pred	1	2	3	4	5	6	7
1 No read	0.88	0.04	0.02	0.02	0.01	0.03	0.00
2 Walk	0.09	0.85	0.04	0.01	0.00	0.00	0.01
3 Out loud	0.13	0.02	0.64	0.17	0.02	0.01	0.01
4 Engaged	0.14	0.02	0.06	0.54	0.12	0.01	0.11
5 Scan	0.08	0.01	0.03	0.39	0.41	0.00	0.08
6 Write/type	0.49	0.01	0.03	0.02	0.05	0.39	0.01
7 Skim	0.13	0.04	0.05	0.47	0.15	0.00	0.16

Table 5: **Reading mode classification** using Gaze, RGB and IMU.

GT \Pred	1	2	3	4
1 No read	0.77	0.07	0.04	0.12
2 Print	0.07	0.55	0.29	0.09
3 Digital	0.08	0.32	0.49	0.11
4 Objects	0.13	0.28	0.30	0.29

(i) Gaze-only

1	2	3	4
0.83	0.04	0.03	0.10
0.08	0.53	0.25	0.14
0.07	0.27	0.53	0.13
0.13	0.18	0.22	0.47

(ii) Gaze+IMU

Table 6: **Reading medium classification** using (i) gaze only (ii) gaze and IMU.

Practical deployment. We also investigate whether model of such size can be run practically. From parallel comparisons, the model can indeed comfortably run real-time on Aria Gen 2 glasses on-device, without the need to off-load the model to online computation. Given the estimated power consumption, the glasses can run for at least 4 hours continuously (inclusive of the base power consumption for basic computation, power delivery and sensor suites, and the thermal constraints). As the model runs on-device without having to send the model input and output back and forth to the server (as would have been done with, say, VLMs), the latency is negligible.

6.5 Ablation studies

Table 4 summarizes our results for ablation studies.

Gaze representation. The gaze processing pipeline involves transforming the retina images into ray angles for each eye, the intersection of which is the 3D gaze point in space, then projecting it onto the 2D image plane. We experiment using all these representations, and find that 3D gaze yields superior results, and pre-differentiating the input with respect to time leads to better generalization.

Head pose representation. With SLAM camaras, we can calculate the visual-intertial odometry (VIO) outputs using visual SLAM, which yields slightly better results compared to raw IMU sensors.

Input frequency and duration. We experiment with varying frequency and duration for eye gaze. We find that higher frequency results in better performance, but also comes with compute tradeoffs. We notice similar trends for IMU.

RGB crop size. While we know that human fovea only covers 2°, we find that a larger crop provides context and covers for errors in gaze estimation. However, the compute also grows quadratically.

Model size. We experimented other model sizes, with XS, S, M, L having 8, 16, 32, and 64 latent dimensions respectively. We also experimented with a pretrained image encoder (MobileNetV3-S) in the XL variant. We find that stronger model results in better results, and notably the S model performs surprisingly well with only 6k parameters.

Robustness to eye tracking precision. While our model is robust to fixed gaze offsets as we only use relative positions, noisy gaze predictions can ruin the gaze pattern. We test for the robustness to noise using our gaze-only model by adding Gaussian noise to the gaze inputs in two settings (i) only at test time and (ii) both during training (as augmentation) and testing. Our results in Figure 8 show the performance degrades with noise, and training with noise helps with robustness.

6.6 Extension: understanding types of reading

Many existing cognitive studies try to understand how humans read, as it is related to understanding human behavior, comprehension, and health. As mentioned, current experiments and datasets are unrepresentative of how we read. In contrast, our dataset extends to "in the wild" settings, and we

hope that our dataset will be useful in advancing the understanding of reading in the real world. Note that we use the same settings as previous experiments (2s time window), which may be limited in such fine-grained classification tasks.

Reading mode classification. Many studies are interested in how people read [3, 42, 28]. We conduct similar studies using our dataset. Specifically, we treat this as a 7-way classification problem (not reading, reading while walking, reading out loud, engaged reading, scanning, reading while writing/typing, skimming), and train the model for this task on our dataset. As shown in Table 5, we find that walking is an obvious category to detect (perhaps due to IMU), followed by reading out loud. Distinguishing between skimming, scanning, and engaged reading proved to be difficult.

Reading medium classification. Inspired by [25] that tries to answer "what" someone is reading, we also conduct similar experiments. In this case, we do not use RGB as the solution would have been trivial, and use the model to classify between four classes (not reading, print media, digital media, objects). We find that the task is difficult, and IMU helps in this case, as shown in Table 6.

7 Broader Impact

Always-on smart glasses raise important questions about social acceptability, both for the wearer and for the public, especially when such technologies are deployed at scale. We hope that the ideas presented in this paper can also help mitigate such concerns.

Safety. Sensitive personal data, such as eye gaze, introduces unique risks. Our algorithm runs fully on-device, which ensures that sensitive information does not need to leave the user's device. This is a step toward stronger privacy protections for wearers. At the same time, we acknowledge that using eye gaze as a signal creates new challenges. Eye movement can reveal intentions, interests, and even emotional states, which raises a distinct category of privacy concerns.

Surveillance. Our work aims to reduce reliance on invasive sensing. First, by leveraging eye gaze, we minimize the required front-camera capture to a very small patch (0.2% of the full image) rather than recording the entire scene. Second, our approach can operate solely on eye gaze data without requiring any camera input. More broadly, eye gaze offers a powerful cue about where the user is looking, which enables RGB capture to be more targeted. This reduces the risk of collecting unintended or intrusive information about bystanders. We hope that future algorithms continue in this direction.

Data Governance. We follow the Project Aria Research Guidelines and will release our system with a Responsible Use Policy to promote ethical research practices and to support safe deployment.

8 Conclusion

Motivated by use cases in contextual AI and other applications, we explore the problem of reading recognition in real-world scenarios, and present a dataset that reflects this nature. We then present a method to solve the task using three modalities, and extend the studies towards reading understanding tasks. There are vast opportunities for future work. Our dataset can be used to study the reading behavior of people in realistic settings in greater detail which links to cognitive understanding. Our proposed protocol allows for scalable future data collection using smart glasses. Additionally, model personalization to address variations in reading speed and style, along with predicting optimal modality activation for enhanced efficiency, represents another promising area for future work.

Acknowledgements

We thank Rowan Postyeni for assistance with data collection, and Michael Goesele, Laurynas Karazija, and Lingni Ma for helpful feedback.

References

[1] Seoyoung Ahn, Conor Kelton, Aruna Balasubramanian, and Greg Zelinsky. Towards predicting reading comprehension from gaze behavior. In *ETRA*, 2020.

[2] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH*, 2023.

[3] Ralf Biedert, Jörn Hees, Andreas Dengel, and Georg Buscher. A robust realtime reading-skimming classifier. In *ETRA*, 2012.

[4] Andreas Bulling, Jamie A Ward, Hans Gellersen, and Gerhard Tröster. Robust recognition of reading activity in transit using wearable electrooculography. In *Pervasive Computing: 6th International Conference, Pervasive 2008 Sydney, Australia, May 19-22, 2008 Proceedings 6*, pages 19–37. Springer, 2008.

[5] Simona Caldani, Christophe-Loïc Gerard, Hugo Peyre, and Maria Pia Bucci. Visual attentional training improves reading capabilities in children with dyslexia: An eye tracker study during a reading task. *Brain sciences*, 10(8), 2020.

[6] Christopher S Campbell and Paul P Maglio. A robust algorithm for reading detection. In *PUI*, 2001.

[7] Diana Castilla, Omar Del Tejo Catalá, Patricia Pons, François Signol, Beatriz Rey, Carlos Suso-Ribera, and Juan-Carlos Perez-Cortes. Improving the understanding of web user behaviors through machine learning analysis of eye-tracking data. *User Modeling and User-Adapted Interaction*, 34(2), 2024.

[8] Xianyu Chen, Ming Jiang, and Qi Zhao. Gazexplain: Learning to predict natural language explanations of visual scanpaths. In *ECCV*, 2024.

[9] Leana Copeland, Tom Gedeon, and B Sumudu U Mendis. Predicting reading comprehension scores from eye movements using artificial neural networks and fuzzy output error. *Artif. Intell. Res.*, 3(3), 2014.

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018.

[11] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, Dan Barnes, Daniel DeTone, David Caruso, Derek Valleroy, Dinesh Ginjupalli, Duncan Frost, Edward Miller, Elias Mueggler, Evgeniy Oleinik, Fan Zhang, Guruprasad Somasundaram, Gustavo Solaira, Harry Lanaras, Henry Howard-Jenkins, Huixuan Tang, Hyo Jin Kim, Jaime Rivera, Ji Luo, Jing Dong, Julian Straub, Kevin Bailey, Kevin Eckenhoff, Lingni Ma, Luis Pesqueira, Mark Schwesinger, Maurizio Monge, Nan Yang, Nick Charron, Nikhil Raina, Omkar Parkhi, Peter Borschowa, Pierre Moulon, Prince Gupta, Raul Mur-Artal, Robbie Pennington, Sachin Kulkarni, Sagar Miglani, Santosh Gondi, Saransh Solanki, Sean Diener, Shangyi Cheng, Simon Green, Steve Saarinen, Suvam Patra, Tassos Mourikis, Thomas Whelan, Tripti Singh, Vasileios Balntas, Vijay Baiyya, Wilson Dreewes, Xiaqing Pan, Yang Lou, Yipu Zhao, Yusuf Mansour, Yuyang Zou, Zhaoyang Lv, Zijian Wang, Mingfei Yan, Carl Ren, Renzo De Nardi, and Richard Newcombe. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023.

[12] Yini Fang, Jingling Yu, Haozheng Zhang, Ralf van der Lans, and Bertram Shi. Oat: Object-level attention transformer for gaze scanpath prediction. In *ECCV*, 2024.

[13] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David

Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022.

[14] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *CVPR*, 2024.

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

[16] Md Farhadul Islam, Meem Arafat Manab, Joyanta Jyoti Mondal, Sarah Zabeen, Fardin Bin Rahman, Md Zahidul Hasan, Farig Sadeque, and Jannatun Noor. Involution fused convnet for classifying eye-tracking patterns of children with autism spectrum disorder. *Engineering Applications of Artificial Intelligence*, 2025.

[17] Soumya Shamarao Jahagirdar, Ajoy Mondal, Yuheng Ren, Omkar M Parkhi, and CV Jawahar. Icdar 2024 competition on reading documents through aria glasses. In *ICDAR*, 2024.

[18] Yang Jin, Lei Zhang, Shi Yan, Bin Fan, and Binglu Wang. Boosting gaze object prediction via pixel-level supervision from vision foundation model. In *ECCV*, 2024.

[19] Marcel A Just and Patricia A Carpenter. A theory of reading: from eye fixations to comprehension. *Psychological review*, 87(4), 1980.

[20] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *ICCV*, 2019.

[21] Conor Kelton, Zijun Wei, Seoyoung Ahn, Aruna Balasubramanian, Samir R Das, Dimitris Samaras, and Gregory Zelinsky. Reading detection in real-time. In *ETRA*, 2019.

[22] Yunsoo Kim, Jinge Wu, Yusuf Abdulle, Yue Gao, and Honghan Wu. Enhancing human-computer interaction in chest x-ray analysis using vision and language model with eye gaze patterns. In *MICCAI*, 2024.

[23] Yan Kong, Sheng Wang, Jiangdong Cai, Zihao Zhao, Zhenrong Shen, Yonghao Li, Manman Fei, and Qian Wang. Gaze-detr: Using expert gaze to reduce false positives in vulvovaginal candidiasis screening. In *MICCAI*, 2024.

[24] Robert Konrad, Nitish Padmanaban, J Gabriel Buckmaster, Kevin C Boyle, and Gordon Wetzstein. Gazegpt: Augmenting human capabilities using gaze-contingent contextual ai for smart eyewear. *arXiv preprint arXiv:2401.17217*, 2024.

[25] Kai Kunze, Yuzuko Utsumi, Yuki Shiga, Koichi Kise, and Andreas Bulling. I know what you are reading: recognition of document types using mobile eye tracking. In *ISWC*, 2013.

[26] Manuel Landsmann, Olivier Augereau, and Koichi Kise. Classification of reading and not reading behavior based on eye movement analysis. In *ISWC*, 2019.

[27] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *ECCV*, 2018.

[28] Wen-Hung Liao, Chin-Wen Chang, and Yi-Chieh Wu. Classification of reading patterns based on gaze information. In *2017 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2017.

[29] Zhi-Yi Lin, Jouh Yeong Chew, Jan van Gemert, and Xucong Zhang. Gazehta: End-to-end gaze target detection with head-target association. *arXiv preprint arXiv:2404.10718*, 2024.

[30] Shaonan Liu, Wenting Chen, Jie Liu, Xiaoling Luo, and Linlin Shen. Gem: Context-aware gaze estimation with visual search behavior matching for chest radiograph. In *MICCAI*, 2024.

[31] Diane C Mézière, Lili Yu, Erik D Reichle, Titus Von Der Malsburg, and Genevieve McArthur. Using eye-tracking measures to predict reading comprehension. *Reading Research Quarterly*, 58(3), 2023.

[32] Sounak Mondal, Seoyoung Ahn, Zhibo Yang, Niranjan Balasubramanian, Dimitris Samaras, Gregory Zelinsky, and Minh Hoai. Look hear: Gaze prediction for speech-directed human attention. In *ECCV*, 2024.

[33] Robert O'Shea. Thumb's rule tested: Visual angle of thumb's width is about 2 deg. *Perception*, 20, 1991.

[34] Süleyman Özdel, Yao Rong, Berat Mert Albaba, Yen-Ling Kuo, Xi Wang, and Enkelejda Kasneci. A transformer-based model for the prediction of human gaze behavior on videos. In *ETRA*, 2024.

[35] Nikhil Raina, Guruprasad Somasundaram, Kang Zheng, Sagar Miglani, Steve Saarinen, Jeff Meissner, Mark Schwesinger, Luis Pesqueira, Ishita Prasad, Edward Miller, Prince Gupta, Mingfei Yan, Richard Newcombe, Carl Ren, and Omkar M Parkhi. Egoblur: Responsible innovation in aria, 2023.

[36] Karolina Rataj, Anna Przekoracka-Krawczyk, and Rob HJ Van der Lubbe. On understanding creative language: the late positive complex and novel metaphor comprehension. *Brain research*, 1678, 2018.

[37] Florian Strohm, Mihai Bâce, and Andreas Bulling. Learning user embeddings from human gaze for personalised saliency prediction. In *ETRA*, 2024.

[38] Enkeleda Thaqi, Mohamed Omar Mantawy, and Enkelejda Kasneci. Sara: Smart ai reading assistant for reading comprehension. In *ETRA*, 2024.

[39] Jie Tian, Lingxiao Yang, Ran Ji, Yuexin Ma, Lan Xu, Jingyi Yu, Ye Shi, and Jingya Wang. Gaze-guided hand-object interaction synthesis: Benchmark and method. *arXiv preprint arXiv:2403.16169*, 2024.

[40] Ru Wang, Zach Potter, Yun Ho, Daniel Killough, Linxiu Zeng, Sanbrita Mondal, and Yuhang Zhao. Gazeprompt: Enhancing low vision people's reading experience with gaze-aware augmentations. In *CHI Conference on Human Factors in Computing Systems*, 2024.

[41] Shaoxuan Wu, Xiao Zhang, Bin Wang, Zhuo Jin, Hansheng Li, and Jun Feng. Gaze-directed vision gnn for mitigating shortcut learning in medical image. In *MICCAI*, 2024.

[42] Keyi Yu, Yang Liu, Alexander G Schwing, and Jian Peng. Fast and accurate text classification: Skimming, rereading and early stopping. In *ICLR*, 2018.

[43] Francesca Zermiani, Prajit Dhar, Ekta Sood, Fabian Kögel, Andreas Bulling, and Maria Wirzberger. Interead: An eye tracking dataset of interrupted reading. In *LREC-COLING*, 2024.

[44] Zehua Zhang, David Crandall, Michael Proulx, Sachin Talathi, and Abhishek Sharma. Can gaze inform egocentric action recognition? In *ETRA*, 2022.

[45] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *CVPR*, 2023.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The introduction makes four claims. The abstract summarizes these claims and they are all prominently present in the paper in that order.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: While not a separate subsection, the limitations were thoroughly discussed including in the dataset (where the training set only contains English language), modalities (that IMU's performance alone is limited), extensions (that the performance on fine-grained classification is limited). We also provide several experiments to test on robustness (Fig. 8) and generalizability (Sec 6.3) which reveals insights on where the model performs less well.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: No theoretical results

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

Justification: As per computer vision models, it is difficult to release every single setting and parameter, but every effort has been made to include implementation details (as in Sec. 5). However we note that the main contribution is not the algorithm or new model architecture. Also, code, model and data will also be released.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will release code, data, and models to the public.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Provided under experimental details, and will be in code release.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Not common in vision tasks as std is small, though main experiments are averaged over 3 runs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Provided in Section 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Nothing to flag

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Included positive impacts, we do not see obvious negative impact of this new task/dataset, but will welcome discussions.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Not high risk

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: Dataset collected ourselves.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Assets will be released with proper guidelines and documentations.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: Human subjects. Instructions included in Supplementary. Exact compensation is not included but we confirm is above minimum wage (USA).

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: IRB approved

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: Not used.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.