
Reading Recognition in the Wild

— *Supplementary Material* —

Charig Yang^{1,2}, Samiul Alam³, Shakhrul Iman Siam³, Michael J. Proulx¹, Lambert Mathias¹,
Kiran Somasundaram¹, Luis Pesqueira¹, James Fort¹, Sheroze Sherifdeen¹, Omkar Parkhi¹,
Carl Ren¹, Mi Zhang³, Yuning Chai¹, Richard Newcombe¹, Hyo Jin Kim¹

¹Meta Reality Labs Research ²VGG, University of Oxford ³The Ohio State University

A Introduction

Additional dataset details. Our dataset is the first instance of reading activity recognition dataset in unconstrained environments and is also the first egocentric dataset with high-frequency eye-tracking (60 Hz) collected through Project Aria. By open-sourcing this dataset, we aim to foster greater involvement from the research community in leveraging the eye gaze data from egocentric devices. Table 1 describes the dataset content in greater detail. In Sections B and C, we provide detailed information on the Seattle and Columbus subsets of the datasets, respectively.

Additional experimental results. In Section D, we provide additional zero-shot experiment results on Columbus subset, which includes ablations and qualitative results. In Section E, we compare our method with other existing solutions, such as VLMs, action recognition models, and alternative baselines. Lastly, we include a discussion in Section F to expand on some experiments, limitations, and future work.

Category	Sub-category	Description/examples
Medium	Print	Books, magazines, newspapers, fliers/brochures
	Digital	Content: news, emails, wikipedia articles, blog posts/forum, e-books, social media, research paper Devices: phone, laptop screen, monitor screen, tablet
	Objects	Nutrition labels, product labels, posters/bulletin board, signs, sticky notes, text on whiteboard
Text type	Paragraphs/Short texts	See Fig. 3 (Main Paper)
	Short texts	
	Non-texts	Children’s books, comic books, instruction manuals, maps, music sheets
	Dynamic texts	Participants are asked to read the subtitles as a video is playing
Multi-task	None	N/A
	While walking	Participants are asked to hold the reading material in their hand(s) while reading and walking
	While writing/typing	Participants are asked to read as they write or type
Mode	Engaged	
	Skimming	Participant is asked to skim the text quickly to get the understand the gist of the text
	Scanning	Participant is given a pre-reading question, and is asked to look at the reading material to find the answer
	Reading out loud	Participant is asked to read the text out loud
Not reading	Daily activities	Five categories: physical exercise, outdoor activity, creative arts, culinary activity, household chores
	Hard negatives	Text is present in scene but user is not reading
Mixed	Alternating sequences	Alternating between reading and not reading, with start/end timestamps
	Mirror setups	A pair of sequences with the same settings, one is reading and another is not reading

Table 1: **Dataset glossary.** This table describes the samples in the dataset in greater detail.

B Reading in the Wild - Seattle Subset

The Seattle subset contains about 80 hours of Aria recordings (total of 1,061 videos) with eye gaze calibration, with 81 participants, taken both indoors and outdoors.

B.1 Type of reading materials covered

The dataset encompasses a wide range of reading materials and content on various mediums. These include:

- **Print:** Books, magazines, newspapers, flyers, and brochures.
- **Digital:** News articles, emails, Wikis, blog posts, bulletin boards, e-books, social networks, and research articles, accessed through four types of digital devices: phones, laptops, monitors, and tablets.
- **Object:** Posters, nutrition/product labels, whiteboards, and sticky notes.

In addition, the dataset includes comic books, maps, video captions, and music sheets, as described in Table 1. We provide the statistics in Figure 1.

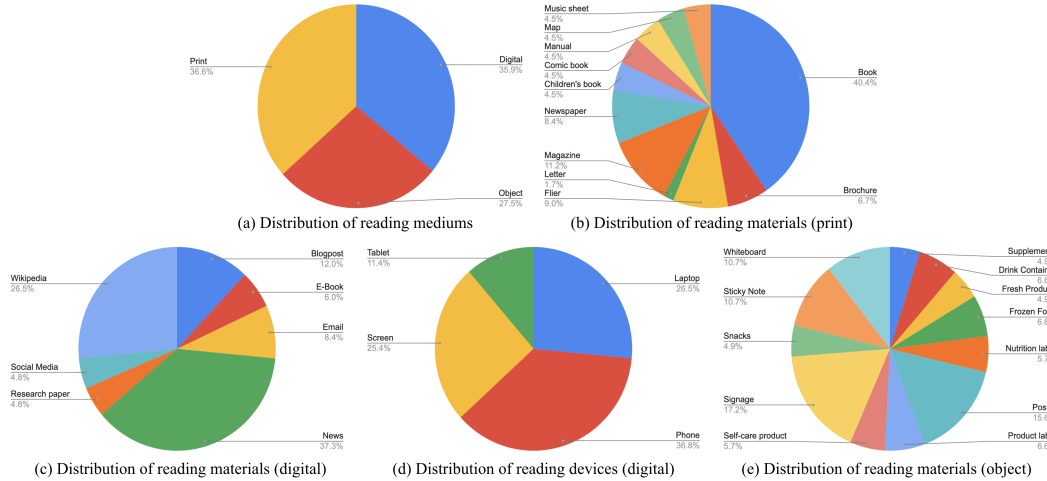


Figure 1: **Distribution of reading materials in Seattle subset.** In (a), we show the distribution of reading mediums. Within each medium (Print, Digital and Object), we then break down the reading materials in (b), (c), and (e). Additionally, for Digital media, we break down by the device involved in the recording. Refer to Fig. 3 (Main Paper) for illustrations.

B.2 Type of reading modes covered

We collected data for different types of reading modes: read out loud, read as you write or type, read as you walk, scanning, and skimming. This is shown in Table 2.

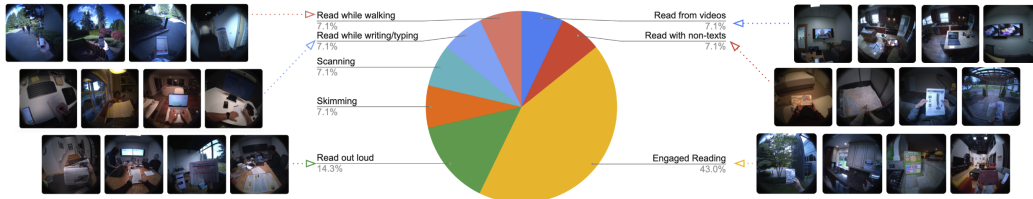
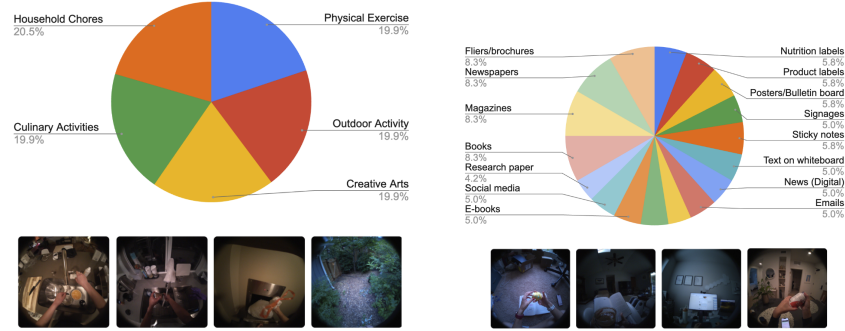


Figure 2: **Distribution of reading modes in Seattle subset** together with illustrations. Almost half of our reading samples is engaged reading, while other scenarios diversely reflect how people read in different scenarios.

B.3 Negative data

We collected two types of negative data: (1) everyday activities that do not involve reading and (2) hard negatives where text is visible but not being read by the participant. Examples include moving

around an object with text (e.g., a vitamin container) without reading it, and spinning a pen while reading material is present but focusing on the pen instead. Metadata is also included to indicate whether the user accidentally read the text. The examples are shown in Figure 3.



(a) Distribution of daily activities in normal negative data (b) Distribution of reading materials in hard negative data

Figure 3: **Distribution of negative data in Seattle subset.** In addition to daily activities, our dataset also includes hard negatives, where the user has a text in scene but is not reading, making it indistinguishable using the RGB stream alone.

B.4 Alternating sequences

We also collected test sequences that alternate between reading and non-reading activities, allowing for the evaluation of temporal localization and latency. We provide the scenario list in Table 2.

#	Summary	Scenario
1	reading + other activities (getting up)	You are reading [Material], then you stop and get up and (go for a walk/prepare some coffee/adjust the lights/ open or close the door or window/ wash your hands/ do any random activity), then you come back to continue reading
2	reading + mind wondering (while still sitting)	You are reading [Material], then while reading your mind wonders, so you stop reading while still holding the material/keeping the same posture (mind wondering), maybe looking around to rest your eyes, maybe staring at nothing, then you get back to reading again.
3	reading text + looking at images	You are reading [Material] containing texts and images, then while reading you inspect the accompanying graphics instead for a while, then you get back to reading again
4	reading + small physical activity	You are reading [Material], then you realize you're sitting in an uncomfortable position, so you stop to adjust your seating, and/or stretch your legs a bit, or move to a different position, chair, or posture, then you come back to continue reading.
5	reading + eating/drinking	You are reading [Material] while enjoying a cup of tea/coffee/drink/food, everytime you take a sip/bite you stop reading, then come back to reading again
6	walking + stopping at a text	You are walking, then you see a [Material – something fixed, like a sign, or a screen], so you stop to read it, then continue the activity after you finish, maybe you walk back to reread the text or then read another text
7	walking + grab something to read	You are walking, then you see a [Material – something that you can grab], so you grab it either stand while reading or sit down to read it, then you grab something else to read similarly
8	cleaning + reading	You are cleaning a room or organizing a table, then you see a [Material], so you grab and sit down to read it. Maybe you are sorting documents and have to briefly read all of them before knowing where to put each one
9	cooking while reading	You are in the kitchen, while making food, you are also multitasking and reading [Material]
10	assembly while reading	You are assembling something or doing something that requires following step by step (text) instructions, periodically consulting the instruction manual, then continuing with the hands-on work.

Table 2: **List of scenarios for alternating sequences.** We collect 12 sequences for each scenario with varied media types, for a total of 120 videos, each with varied reading materials. We also provide annotated timestamps for each start/stop section.

B.5 Demographics

The demographics are presented in Figure 4. We ensure a diverse range of demographics across gender and age in our collection.

B.6 Metadata

The metadata provides valuable information about the reading task and the associated reading material, reading medium, summaries, and multiple-choice question IDs and answers if applicable. This is

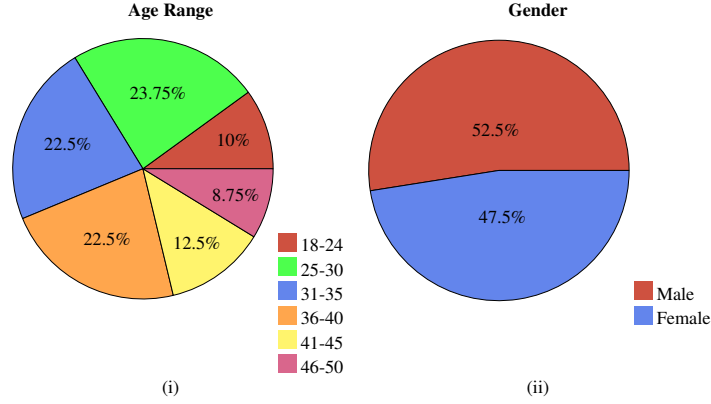


Figure 4: **Demographic statistics of the Seattle subset** (i) shows the age group, (ii) shows gender distribution

shown in Figure 5. In addition to these, we also collect personal demographics (age range, gender) as part of the metadata.

```

"name": "13. Write or type texts - Read Out Loud 21",
"notes": "I wrote about tattoos, the week, what food I want.",
"recording_profile": "profile28",
"tags": [
  {
    "title": "How long is the text?",
    "tags": [
      "Equal or greater than 4 sentences"
    ]
  },
  {
    "title": "Where did you write/type texts? ",
    "tags": [
      "Phone"
    ]
  },
  {
    "title": "How did you write the text? ",
    "tags": [
      "Typing"
    ]
  }
]

```

Figure 5: **Example of metadata for Seattle subset.** The metadata contains several useful information to facilitate further research, including both multiple-choice questions and short answers.

B.7 Protocols

A successful data collection protocol ensures efficient collection while guaranteeing the quality. Reading is a complex process that includes word recognition, which encompasses visual processing and language decoding, and comprehension, which involves linking information to memory and cognitive processing. Therefore, using a single protocol often proves challenging. For example, asking the participants to read text out loud guarantees that every single word is read / seen. However, it cannot prevent participants' minds from wandering. On the other hand, giving the participants a set of questions to answer can make sure that the participants actually understood the text. However, it makes the task harder than necessary as it also puts the subjects' other cognitive capabilities (e.g. memory) in question.

Instead, different protocols can be tailored to specific reading modes and materials. To address various reading modes such as scanning, skimming, or engaged reading, and to accommodate different types

of reading materials, we use the following strategies. Please note that this captures only a portion of the entire problem space:

- **Pre-reading questions:** Presenting questions before the reading task can focus participants' attention on specific details, making it effective for scanning tasks where the goal is to quickly locate particular information.
- **Post-reading questions and summaries:** Activities such as answering questions, summarizing the content (which can be evaluated using LLMs) can help assess comprehension by indicating whether participants have understood and can apply the information they've read. Specifically, we ask the participant to summarize what they read in a few words for all our tasks. For a set of reading materials (12 digital contents), we prepared multiple-choice questions and had the participant answer them.
- **Verbalization and recall:** For shorter texts, asking participants to read aloud the text or recalling the text is an effective way to ensure that person read the text. Reading out loud is also effective for reading while writing or typing, as the pace of writing is generally slower than speaking. By asking the participant to read out loud can also capture valuable insights into how reading behavior changes between reading silently and reading aloud, especially since the pace of speaking can influence gaze behavior.
- **Embedded tags:** Techniques like embedding AprilTags within the reading material can provide additional signals on whether the participant comprehended the content (For example, if the instruction is to "Go to page 20 if you decide to escape," and the participant turns to page 20, it indicates that they paid attention to content.). This approach also enables automatic segmentation of reading activity (starting and end times). We prepared 10 CYOA (Choose Your Own Adventures) books with AprilTags appended to specific pages: April Tag 0 for the first page after the cover, indicating the start of reading; April Tag 1 for the branching page; April Tag 2 on the page where they should land if they correctly read the instructions.

In addition, we also tailored protocols for specific reading materials with non-texts:

- **Music sheets:** We pre-screen for participants who can read music notes, and they are asked to hum the notes as they read.
- **Maps:** Participants are asked to describe how to navigate from place A to place B as they read the map.
- **Instruction manuals:** We employ two approaches: participants are asked to summarize the steps and also to follow the instructions to assemble or set up items like board games, Lego-alikes, and origami.
- **Video captions:** We turned off the audio, such that the participants are forced to read captions. They are also asked to provide the summary of the video.

B.7.1 Scalable first-person annotation

To avoid manual labeling of timestamps, we instruct the participants to say "start reading!" whenever they start reading, and "finished reading!" whenever they finish. This approach allows us to use a speech recognition model WhisperX [1] to obtain accurate timestamps without requiring manual annotations, making the process scalable. It is also important to have the participant to annotate in this manner, as it would be challenging for a third party to accurately segment the reading sessions.

B.8 Annotation

The majority of our data is automatically labeled using WhisperX (speech recognition) and AprilTag detection. However, for certain tasks, such as our test sequences, where participants switch between reading and non-reading activities, manual inspection and re-labeling are necessary to ensure accuracy. For reading out loud tasks, we use the beginning and end of speech to mark the reading portion. For silent reading tasks without AprilTags, we use a simple verbal cue ("start/finished reading") to mark the beginning and end of each reading session.

C Reading in the Wild - Columbus Subset

The Columbus subset contains data collected from 30 subjects. The data was collected using the Aria wearable glasses indoors on the university campus. The study was reviewed and approved by the university’s Institutional Review Board (IRB). Similar to the Seattle subset, the raw data include a single RGB (30Hz 1408p, 110 FoV°), two SLAM(150°FoV) and two infrared eye tracking (60Hz, calibrated) video streams, and two IMU data streams. Additionally, we collected spatial audio from the glasses and recorded task completion times of each segment for annotation. The subset contains 640 recordings corresponding to 18 hours of reading/non-reading tasks.

C.1 Type of reading materials covered

The Columbus subset includes a diverse range of reading materials designed to evaluate performance across various contexts. To create a diverse dataset, we considered 3 aspects: *medium*, *content length*, and *content type*. Figure 6 illustrates the distribution of different (a) mediums, (b) content lengths and (c) content types.

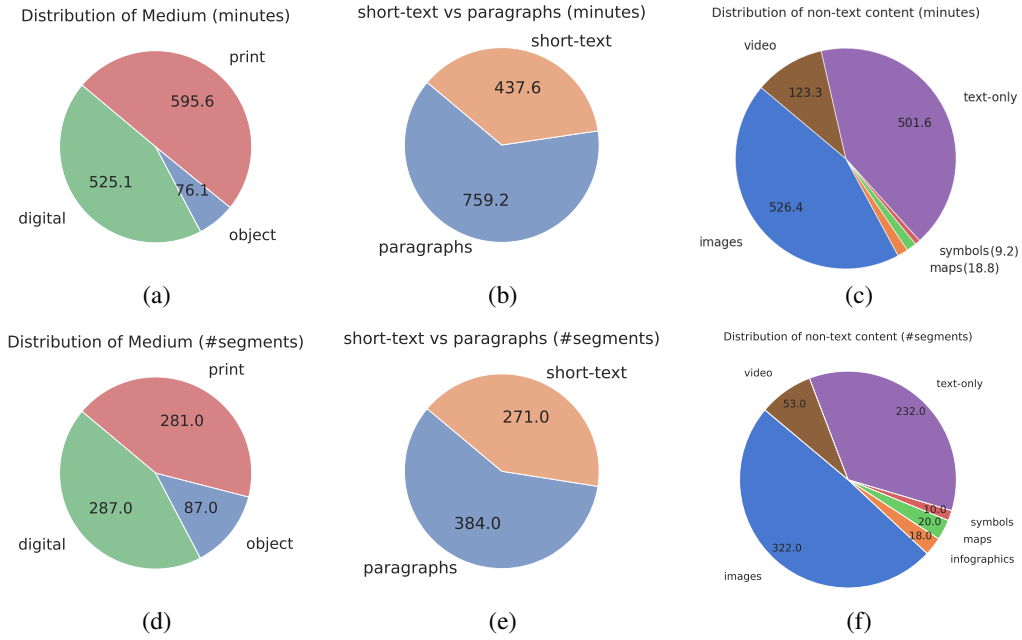


Figure 6: **Types of reading materials covered.** The top row shows the distributions of (a) medium (b) text types (c) non-text content, in minutes. The bottom row shows the same distributions in terms of number of segments.

‘Medium’ indicates what kind of device or object the subject is reading from. The following is a list of reading materials included for different mediums. Similarly to the Seattle subset, we indicate the device (phone, laptop, tablet, screen). For objects, we also indicate the shape of the item which are ‘cylindrical’, ‘flat’, ‘box’, ‘poster’, or ‘other’.

- **Print:** Books, Booklets, Pamphlets, Spreadsheets, Magazines, Flyers, Instruction Manuals, Handwritten Text, Maps.
- **Digital:** Wikis, Blogs, Forums, Research Articles, News Articles, E-Shopping Sites, Video Sites.
- **Object:** Posters, Nutrition/Product Labels, Flashcards, Signs.

‘Content length’ consists of (1) ‘paragraphs’ where the content requires the subject to read one or more sentences in the text. The scanpath for these usually results in smooth horizontal lines; (2) ‘short-text’ where the text is scattered across the page as short texts. These can include image captions, tabular datasheets, location names on maps, etc. The scan path for short-text is therefore more irregular.

‘Content type’ indicates the type of content the user is reading. It can be text-only, where there is only text in the material, or it could be text embedded with other visuals like images, videos, maps, infographics, or symbols.

Additionally, within each medium, the reading material is also varied to create a diverse dataset. The distribution of different platforms for each medium is shown in Figure 7.

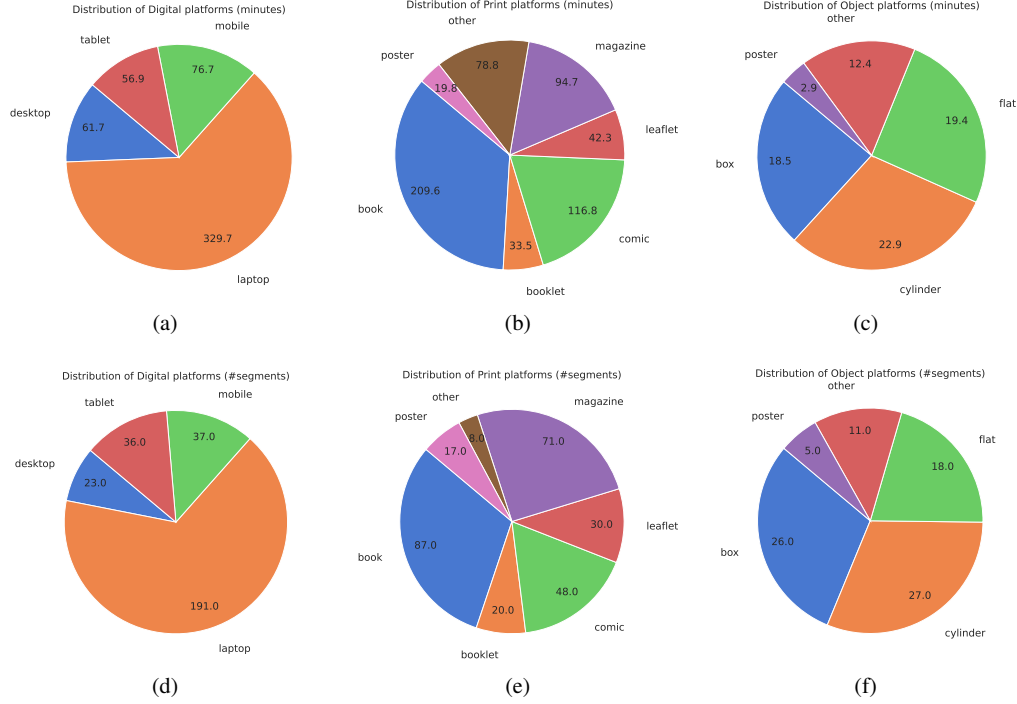


Figure 7: **Platform distributions by medium.** The top row shows the total duration in minutes for digital, print, and object platforms, while the bottom row displays the corresponding number of recordings.

C.2 Type of reading modes covered

In the Columbus subset, we have both instances of subject reading or positive cases and not reading or negative cases. However, similar to the Seattle subset, the mode of reading can be different within positive instances. As such we have additional metadata to capture the reading mode. There are 2 reading modes in the subset; regular engaged reading and scanning. Engaged reading indicates cases where the subject is asked to read a section of text and is expected to answer a question based on the text. The question is revealed only after completing reading. In contrast, when scanning, the subject is given a question and asked to find the answer in a particular section of text. In the Columbus Dataset, we have 30.6 minutes of scanning data and 786.7 minutes of engaged reading which corresponds to 36 recordings and 435 recordings respectively.

C.3 Negative data

Similar to the Seattle Dataset we also collect two types of Negative Data: (1) everyday activities and (2) hard negatives. In everyday activities, subjects are given a general tasks like watching a video or walking in the corridor and asked not to read anything. In this cases there is only minimal or no reading material in the field of view of the subject. For ‘Hard Negatives’, subjects are instructed not to read as well. But in contrast to everyday activity instances, subjects will have significant reading material in their field of view. They are specifically instructed to only focus on non-text content like images. In the data we have 110.4 minutes of everyday activity data and 147 minutes of hard negative data corresponding to 42 and 127 recordings respectively.

C.4 Mirror setups

The Columbus Dataset contains mirror setups where the reading material is the same. The subject performs two tasks - one engaged reading and the other is either not-reading or scanning. Overall we have 28 such sequence pairs with mirror setups in the subset (25 reading and not-reading pairs and 3 reading and scanning pairs). The reading materials included object and print mediums but no digital mediums. The setup is designed to test the model’s discriminative abilities in a symmetrical environment where the primary difference is reading behavior.

Figures 8, 9, and 10 illustrate three distinct mirror setups from the Columbus Dataset, each designed to evaluate the model’s performance in distinguishing between reading and non-reading tasks under controlled conditions. The reading materials remain identical, and variations arise solely from the task being performed. We analyze the performance across 3 frames on Gaze only, RGB Only, Gaze+RGB, and Gaze+RGB+IMU modality combinations. The RGB crop used by the model is shown in the red box on the figures and the gaze pattern is indicated by the connected dots with the red dot being the last and most recent gaze point and blue being the start of the sequence. Note that the gaze pattern fed to the model spans 2s with 120 gaze points in total. However, in the figures only 8 dots are shown which are 250ms apart.

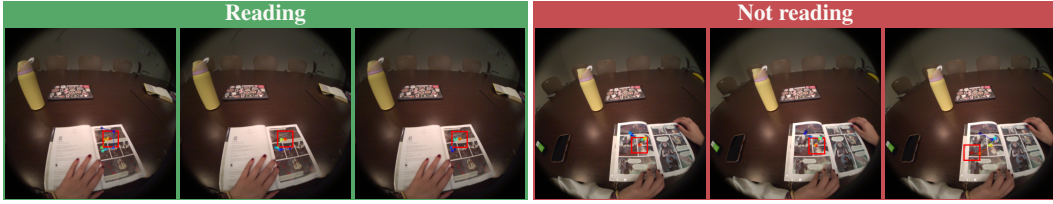


Figure 8: **Mirror Setup: Print Medium.** Here subject is asked to read a comic vs. when asked to not read anything but instead look at pictures only.



Figure 9: **Mirror Setup: Walking in Corridor.** Here the subject is asked to read the room numbers and signs in a corridor vs when asked to traverse through normally.

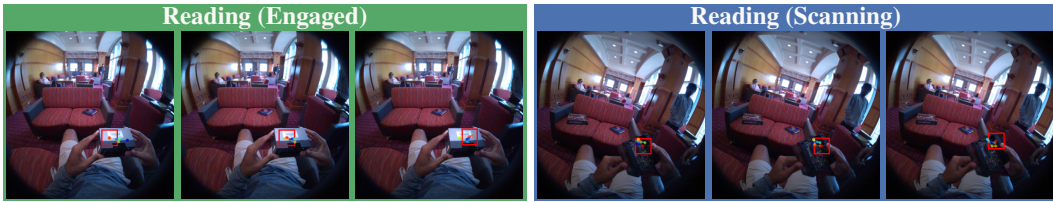


Figure 10: **Mirror Setup: Reading vs Searching.** Here the subject is asked to read the serial numbers in a circuit board vs when asked to specifically search and count the number of resistors.

C.5 Annotation

To annotate the raw data and extract segments, we utilized a labeling tool developed in PyQt5. Figure 11 outlines the graphical user interface of the annotation tool. The annotation process begins by importing the raw video file, which is then manually input the timing information using the timeline slider and 'set' button. For each segment, we record the start and end times in seconds to mark the exact duration the segment appears in the video. To ensure a clean transition between segments, we trim a small portion from the beginning and end of each segment, avoiding any unintended overlap or transition frames. Each segment corresponds to a recording in the subset. To ensure easy and

consistent selection, we use drop-down menus to store the values for ‘Medium’, ‘Content-Type’, ‘Platform’, and ‘Activity Type’. After annotating each video, segment information is saved in a separate CSV file for further processing.

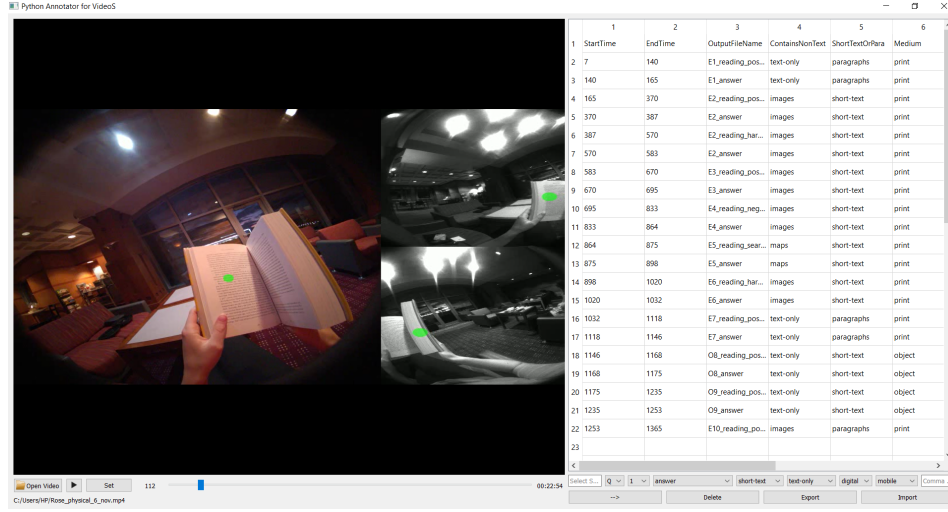


Figure 11: **Graphical interface of annotation tool.** The user interface consists of the RGB video preview with sound and the gaze point overlaid on top shown in green. The interface allows adjusting the start and end time of each recording as well as annotating metadata like content-type, content-length and medium.

C.6 Demographics

We posted flyers in various locations to inform interested people about the study. Participants reached out to us through email and a session was booked for them based on availability. Among the 30 participants, 15 were male and 15 were female. The age of the participants ranged from 18 to 34 with 13 participants between 18-24 years old and 17 between 25-34. 12 participants were native English speakers and 18 were non-native speakers. Participants came from a wide range of backgrounds and had at least a high school degree. To be eligible, participants had to be able to read without glasses. Of the participants, 21 had perfect vision and 7 wore contacts. 2 participants had mild myopia but did not require glasses to read.

C.7 Metadata

The Columbus subset contains metadata related to demographics, tasks, and scenario details.

C.8 Protocols

For the Columbus subset, we follow a single protocol which was approved by IRB. We follow this protocol which involves both pre-session preparations, collection of data and final processing steps. The protocol steps are discussed as follows:

C.8.1 Pre-session Briefing

Before starting session, we brief the subject on the motivation and objective of the study, a description of the kind of reading material we will provide, an overview of all the sessions, and the nature of tasks like reading, hard negatives, mirror setup tasks, etc. We provide a consent form asking the participant to acknowledge consent for the study and their rights as a participant. Once the paperwork is done and the participant is briefed and ready, we begin data collection.

C.8.2 Data collection

The data collection process for this study consisted of three primary sessions, with a total duration of approximately one hour. The initial two sessions (digital medium session and print/object medium session) took place indoors inside a room, while the optional third session was conducted on corridors or balconies if time and participant availability permitted. The collection protocol is discussed below.

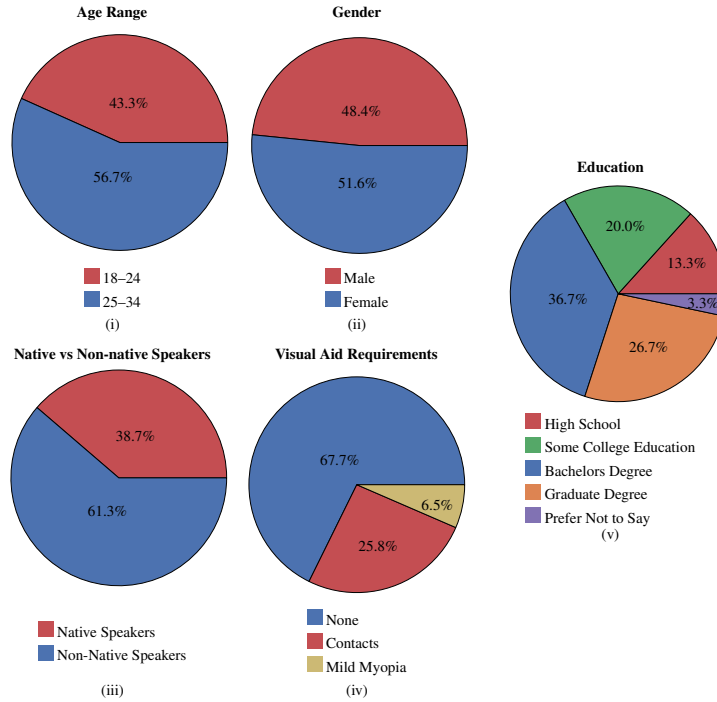


Figure 12: **Demographic statistics of the Columbus subset.** An overview of the demographic distribution is shown: (i) Age Range, (ii) Gender, (iii) Native vs Non-native Speakers, (iv) Visual Aid Requirements, and (v) Education.

Pre-session Preparation The process began with participants filling out a demographic questionnaire in a controlled indoor environment. Following this, an attendant explained how to calibrate the Aria glasses, and once participants were familiar with the procedure, calibration was performed. After this, we proceeded with the first session.

Digital Medium Session. The session generally involved interacting with digital media: participants were directed via a Qualtrics survey to visit specific websites and complete tasks such as reading certain sections, searching for information, or browsing without reading. The tasks varied, and each participant received different instructions. Responses to these tasks were then collected to assess comprehension or adherence to instructions. It is important to note that any recordings where participants inadvertently read text, contrary to instructions, were discarded or trimmed. Throughout the session, each participant used one specific device (laptop, tablet, mobile, or desktop), although they were permitted to bring their own device. The survey recorded the completion time for each question, aiding in the segmentation of the recording for later analysis. The digital medium session was generally conducted first, but in a few cases, we started with the print and object medium session.

Print/Object Medium Session. After a brief rest period of 5-10 minutes, the second session commenced, focusing on print and object mediums. Participants were provided with various printed materials, such as books, magazines, flyers, and instruction manuals, or objects with embedded text like product packaging. They were instructed to read specified sections, and the start and end times of their reading were manually logged by an attendant for segmenting the session later. Subsequently, participants answered questions to verify their engagement with the text.

Impromptu session If there is time remaining, we then arrange the last session. This is a short session in which we ask the participants to do random tasks which include both reading and not reading. These tasks could be reading a wall-mounted poster or flyer, walking along the corridor reading room numbers or signs, or reading a wall-mounted map. Like the previous two sessions, participants may be instructed to either read or not read any text.

```

"name": "<name-of-recording>",
"is-reading": True,
"mode": "scanning",
"negative-type": "N/A"
"scenario": {
  "contains-non-text": True,
  "short-text-or-para": "Paragraphs",
  "platform": "Laptop",
  "mode": "scanning",
  "ExtraTags": [
    "Had contacts",
  ],
},
"subject": {
  "SubjectID": "12345",
  "AgeRange": "25-34",
  "Gender": "Male",
  "EducationLevel": "Graduate",
  "Corrective Lenses": "Contacts",
  "Specialization": "Computer Science",
  "Country": "USA"
}

```

Figure 13: **Example metadata for Columbus subset.** The metadata contained is slightly different than that of the Seattle subset, allowing for different directions to explore for further research.

C.8.3 Post-processing

After the sessions, we generate draft previews of each session for segmentation. The previews are low quality RGB videos of the session with the gaze overlaid on the videos. We get the approximate segment break points from the survey data. For each session we manually adjust the start and end times of each segment. Additionally, we manually annotate medium, platform, content-length, content-type, and language for each segment at this time by looking at the previews. We also add extra tags we think may be of importance. We save the manually adjusted timing data and use it to segment and export the data into individual recordings containing only the relevant data, excluding the answering and instruction phases. We process these recordings using a cloud service to get the eye-tracking annotations and open-loop trajectory based on the SLAM data. We review the generated recordings and annotations and do another pass on the manual annotations to make sure the segmentation and metadata are correct.

C.8.4 De-identification

To conform with University policy and the guidelines approved by the IRB, we are required to de-identify any descriptions or data that may identify the participant. We deidentify any descriptions, text and filenames containing the participant's identifiable data. Despite precautions, some faces of the attendant and bystanders may be present in the video recordings. We use EgoBlur [6] to blur out any such faces that may have been captured by the glasses in the segmented recordings.

D Supplementary Results on Columbus Subset

D.1 Main Results

Gaze	RGB	IMU	Acc	F1	AUC	$P_{R=90.0}$	$T_{R=90.0}$	$Acc_{R=90.0}$	$F1_{R=90.0}$
✓			77.1	84.0	95.3	84.1	29.2	79.1	86.9
	✓		76.7	84.5	92.1	83.4	29.4	78.5	86.6
		✓	71.4	82.2	81.9	78.5	42.3	73.3	83.9
✓		✓	76.7	84.0	94.9	82.7	26.6	77.8	86.2
	✓	✓	77.8	85.6	92.4	83.6	36.9	78.6	86.7
✓	✓		82.8	88.7	96.4	88.2	40.3	83.0	89.1
✓	✓	✓	82.9	88.8	96.4	88.2	42.4	82.9	89.1

Table 3: **Results on the Columbus subset.** The table includes different combinations of these modalities (using only single modality, using any combination of two modalities, and using all modalities). The first four metrics are: accuracy (Acc), F1 score (F1), area under the curve (AUC), and the Precision at Recall of 90% ($P_{R=90.0}$). Consistent with Seattle subset in the main paper, we show the usefulness of combining gaze and RGB. IMU becomes less useful with less daily activities. In the last three columns, the confidence threshold $T_{R=90.0}$ for binary classification is set where Recall is equal to 90%. We report the accuracy metrics at this confidence threshold. We notice lowered threshold to reach such high recall, adversely affecting the precision in the process.

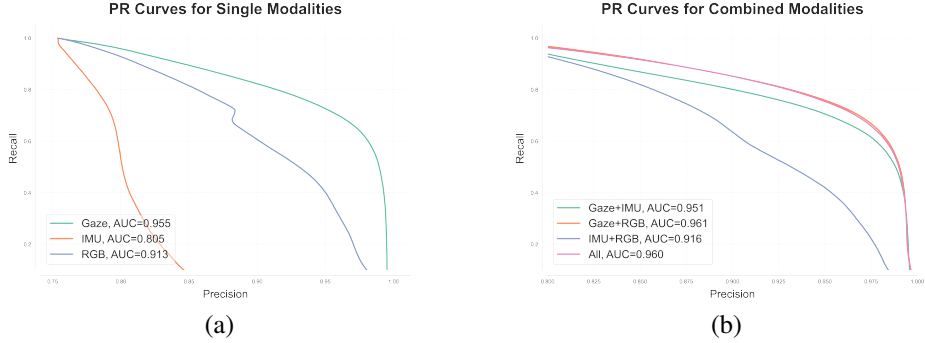


Figure 14: **PR Curves for different modalities** We compare the PR curves of different modalities. (a) shows the curves for individual modalities and (b) compares how combining different modalities influence performance

The results presented in Table 3 highlight the effectiveness of various modalities—Gaze, RGB, and IMU - in detecting reading activity on the Columbus Dataset. We make some key observations:

Single Modality: Firstly, we find that the Gaze alone achieves the highest performance among single-modality setups. This is reasonable as gaze data provides the most discriminative features for detecting reading behavior. Given the larger proportion of hard negatives and smaller daily activities, we observe a drop in performance for RGB and IMU compared to the Seattle subset.

Combined Modalities: Consistent with the Seattle subset, we find that combining different modalities yield better results, with the Gaze+RGB+IMU model giving best performance. However, we notice that the contribution of IMU data is relatively marginal in this subset when gaze and RGB features are already present. This is reasonable as the Columbus subset consists mostly of engaged reading tasks where the subject seldom moves their head.

Confidence threshold tuning: In contextual AI use cases, we expect the model to be prioritizing recall over precision to minimize false negatives while accepting some false positives. Hence, the last three columns of the table presents the threshold-tuned results for Gaze used individually and in combination with RGB and IMU modalities, where the confidence threshold for binary classification is set where recall is equal to 90%. This results in some trade-off in precision in the process.

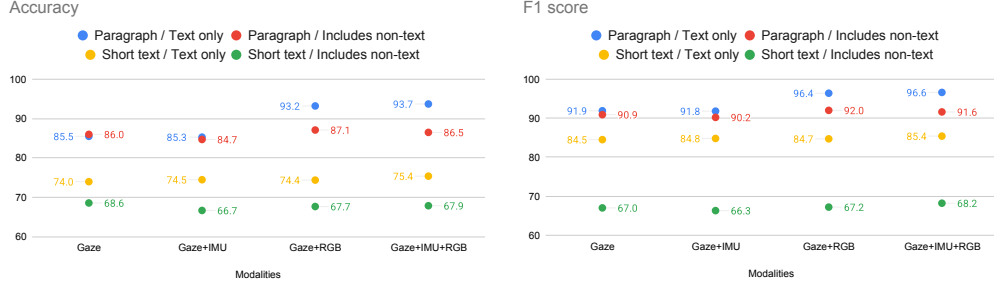


Figure 15: **Breakdown by content length and content type in Columbus subset.** The figures show the Accuracy and F1 scores respectively of different combinations of content length (Paragraph vs Short text) and content type (Text only vs Includes non-text), across different modalities. We show that reading detection works better on paragraphs and text-only cases.

D.2 Result breakdown

Content type and content length. Figure 15 provides a detailed evaluation of the role of gaze as a primary modality for reading detection and how the addition of RGB and IMU data impacts performance across different content types (samples that have text only vs. those with non-text items) and content lengths (paragraphs vs. short texts). We make several observations:

First, in terms of text length, the performance for paragraphs (blue and red) is significantly better than short texts (yellow and green) across modality combinations and performance metrics. The scan pattern is more irregular in short texts making it harder for gaze data to detect reading behavior.

Second, we notice that text-only models performs better than ones that include non-text elements. We notice that this difference is more pronounced in models using RGB. This is sensible, as non-text elements introduce visual distractions, making it harder for the model to focus on reading cues.

Lastly, we notice the trend consistent with previous observations regarding different modality combinations, with all modalities performing best among all the options.

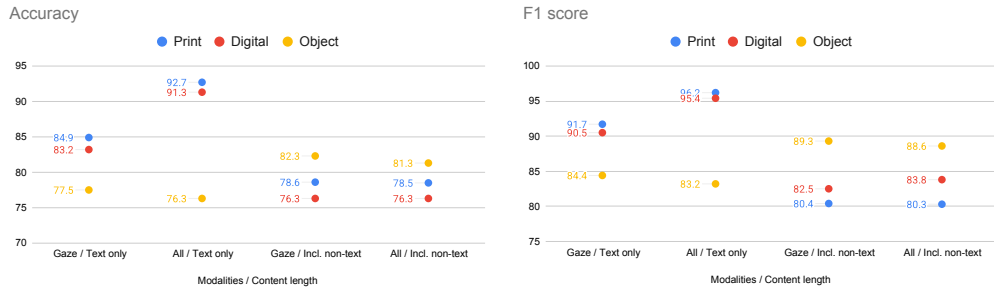


Figure 16: **Breakdown by medium and content type in Columbus subset.** The figures show the Accuracy and F1 scores respectively of different combinations of medium (Print vs Digital vs Object) and content type (Text only vs Includes non-text), across different modalities. We show that reading detection works better on print and digital media.

Medium and content type. Figure 16 evaluates the performance of gaze as the primary modality and the contribution of RGB and IMU data across three mediums (Digital, Object, and Print) and two content types (Text Only and Contains Non-Text). The results reveal insights into the interaction of modality combinations with medium and content type.

For reading with text only, we notice that the performance for print and digital media are higher than that of object. The inverse is true for reading content that contains non-text. We think that this reflects the nature of the real-world reading content in the training data *i.e.* print and digital media tend to be mostly containing texts, and objects tend to include non-text elements.

Another interesting observation is that the performance tends to stagnate for both objects and other mediums including non-text despite more modalities. This may be because of the potential misleading cues provided by the RGB stream in these cases. For text-only print and digital media, the performance sees a significant boost by adding more modalities.

D.3 Cross-Language

Although left-to-right (LTR) is the most common writing direction in modern writing systems and is used in popular languages like English, Hindi, Spanish, and Bengali, there are widely used languages that use alternate writing directions like right-to-left script (RTL) used in Arabic where text is written from right to left and top-to-bottom (TTB) used in Traditional Chinese, Japanese and Korean language where text is written top to bottom and right to left. In the Columbus subset, we include Arabic and Traditional Chinese to analyze the performance of reading detection for RTL and TTB using Arabic and Traditional Chinese respectively.

Table 4 present the performance of the default model on Arabic text under two conditions: (1) the original gaze sequence and (2) the gaze sequence flipped. The results show the variations in performance based on modality combinations.

Augment	Gaze	RGB	IMU	F1	Acc
None	✓			23.8	21.0
	✓	✓		82.5	70.8
	✓	✓	✓	87.5	78.8
Flip	✓			63.8	51.5
	✓	✓		88.4	79.6
	✓	✓	✓	91.5	85.0

Table 4: **Results on Arabic with and without augmentation.** We show that flipping the gaze horizontally allows right-to-left language to be read better. RGB models show good generalization despite only being trained in English.

Augment	Gaze	RGB	IMU	F1	Acc
None	✓			51.6	35.5
	✓	✓		77.4	63.4
	✓	✓	✓	86.4	76.3
Rotate	✓			91.9	85.1
	✓	✓		94.2	89.1
	✓	✓	✓	95.9	92.3

Table 5: **Results on Traditional Chinese with and without augmentation.** We show that rotating the gaze allows top-to-bottom language to be read better. Again, RGB models show good generalization despite only being trained in English.

In Table 4, which evaluates the original gaze sequence, gaze as a single modality achieves the lowest performance (F1 = 23.8%). The performance is drastically affected since Arabic is read from right to left and the model was not trained on such data. We also evaluate performance with the gaze sequence flipped. This is done by flipping the x-axis input. Here, the accuracy on Gaze only improves to 52.5%. Other modalities are less affected by the reading direction. Similarly, models with combined modalities using gaze show a noticeable improvement compared to the unaugmented original gaze sequence. This suggests that augmenting the gaze sequence to reflect the directionality of Arabic text better improves the effectiveness of gaze data in multimodal setups.

Next, Table 5 compares model performance on Traditional Chinese text with no augmentation vs when the gaze is rotated. Similar to Arabic, the gaze struggles to read vertical texts without augmentation (though not as much, as some vertical texts do exist even for English, especially in objects). Still, we notice a significant performance boost when the gaze is rotated. Similarly, rotating the gaze also helps when the gaze is used with other modalities.

D.4 Mirror Setups

The performance for Mirror Setups are shown in Table 6. Compared to the entire Columbus subset, we see a significant drop in performance for RGB modality. This is sensible, as this part of the dataset is designed so that the RGB appears difficult to distinguish. The performance of the other two modalities is similar to that of the whole Columbus dataset, as was expected, resulting in a slightly worse multimodal model when modalities are combined.

We also analyze the results for specific examples introduced in Section C.4.

	Gaze	RGB	IMU	F1	Acc	AUC	P _{R=90.0}
Single	✓			82.4	78.9	92.7	75.7
		✓		69.1	66.1	82.1	67.6
			✓	74.5	64.6	75.8	65.5
Dual	✓		✓	80.8	76.6	91.8	71.9
		✓	✓	73.6	69.2	83.5	69.0
	✓	✓		80.6	77.4	92.1	73.8
All	✓	✓	✓	80.9	77.3	91.8	72.4

Table 6: **Performance of model on mirror setups.** The performance for RGB is significantly lower, while other modalities remain similar.

In Figure 8, the subject is asked to read a comic in the first setup and then look at the pictures in the comic without reading the second setup. The model accurately detects reading and not reading in both setups across all 3 frames using Gaze only, RGB Only, Gaze and RGB, and all modalities.

Figure 9 presents a scenario where the subject is asked to read room numbers and signs in a corridor in the first task and then traverse the corridor without reading in the second task. Here, due to the small size of the reading material and the patterns of the eye gaze, none of the modality combinations taken individually or combined can accurately detect reading in any of the 3 frames. However, Gaze only and RGB and Gaze is able to detect not-reading in the second scenario correctly across all 3 frames. RGB only however detects the second scenario as reading across all 3 frames.

Figure 10 explores a more fine-grained task differentiation where the subject is asked to read serial numbers on a circuit board versus search and count the number of resistors. Before the sessions subjects were briefed on how to identify resistors on the circuit board and what serial numbers to read. Here, all of Gaze only, RGB Only, Gaze+RGB, and Gaze+RGB+IMU successfully detect reading in the reading task, however, RGB Only detects not reading on the search task. The other combinations successfully detect reading on the searching task too.

D.5 Qualitative Results – Partial Success

We visualize some samples in the Columbus subset where different combinations succeed or fail. We note some interesting cases here.

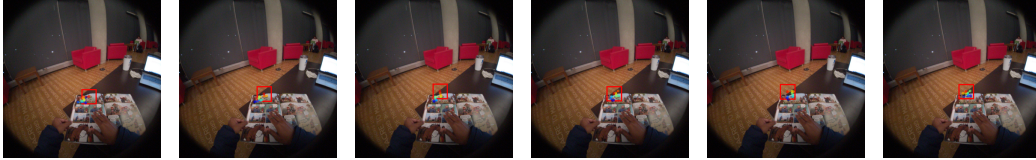


Figure 17: **RGB fails but Gaze succeeds.** Failure case across 6 frames where RGB Fails but Gaze works. Notice that the RGB crop indicated in red has partial coverage of the reading material.

In Figure 17, we observe a case where RGB fails but Gaze succeeds. Here, the RGB crop, marked in red, provides only partial coverage of the reading material, likely leading to misclassification. Despite the suboptimal visual information, Gaze remains effective by directly capturing the subject’s attention and focus. This case underscores the robustness of gaze pattern in reading detection.

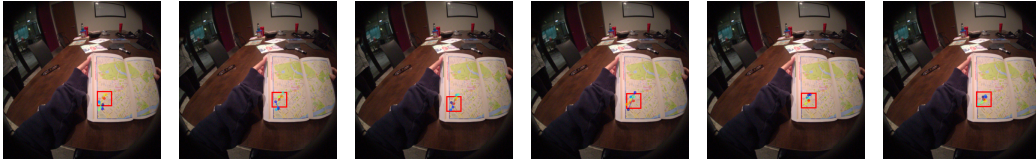


Figure 18: **Gaze fails but RGB succeeds.** Here, the subject is reading a map where text is irregularly placed across the field of view, making gaze patterns more sporadic.

In contrast, Figure 18 presents a case where gaze fails but RGB works effectively. Here, the subject is reading a map. The subject is asked specifically to read all the names on the map along a particular

route. The text being the names of streets and roads along the route does not follow a particular pattern. The irregular placement causes sporadic gaze patterns instead of the horizontal pattern seen in regular reading, making Gaze less reliable for reading detection. RGB, however, can capture the broader visual context of the map and successfully identify the reading behavior. This case along with the previous discussion of Figure 17 highlights the complementary nature of RGB, especially in scenarios with spatially scattered reading content.

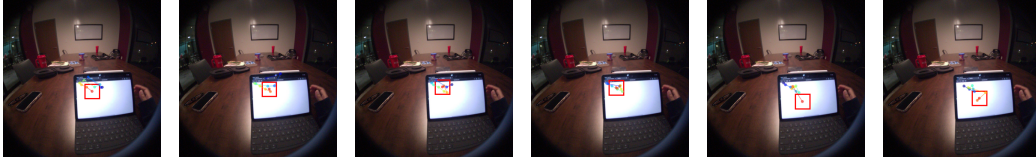


Figure 19: **Individual modality fails, combined modality succeeds.** Here, both gaze and RGB fail individually but succeed when combined.

In Figure 19, both Gaze and RGB alone fail to detect reading behavior, but their combination succeeds. The multimodal setup leverages gaze to pinpoint attention regions and RGB to provide contextual information, leading to accurate detection. Note that we have found no cases in the Columbus subset where both Gaze and RGB individually fail but Gaze and RGB taken together succeed.



Figure 20: **Misleading RGB: Gaze measurement error.** Failure case across 6 frames where Gaze succeeds but RGB Only and Gaze with RGB fail. Note that here the eye gaze is offset due to measurement error, putting the RGB crop outside the reading material.

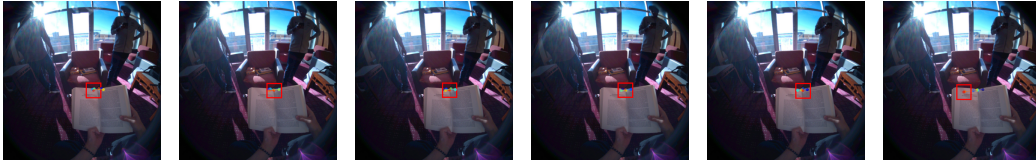


Figure 21: **Misleading RGB: Partial coverage.** Failure case across 6 frames where Gaze succeeds but RGB Only and Gaze with RGB fail. Notice that the RGB crop indicated in red has partial coverage of the reading material.

In Figure 17, Figure 18 and Figure 19, Gaze+RGB successfully detected reading. In contrast, Figures 20 and 21 illustrate cases where Gaze succeeds but RGB, either alone or in combination with Gaze, fails. Figure 20 shows a measurement error where the gaze point is outside the reading material and so the RGB crop provides inaccurate contextual coverage, leading to misclassification. In Figure 21, the subject is reading at the top edge of a book and so the RGB crop is partially on the reading material. This is similar to the case in Figure 17 except, here, the RGB+Gaze combination fails.

D.6 Qualitative Results – Failure Cases

Figures 22 and 23 depict complete failure cases where all modalities and their combinations fail to detect reading behavior. Figure 22 is an extreme case of Figure 18, where the subject is searching for a particular location in a map instead of reading individual street or road names along a route. Figure 23 highlights the challenge of detecting reading behavior while walking and reading room numbers. Note that although the model was trained on sequences where the subject was reading something in their hands while walking, it was not trained on reading stationary objects while walking, which potentially adds to the difficulty.

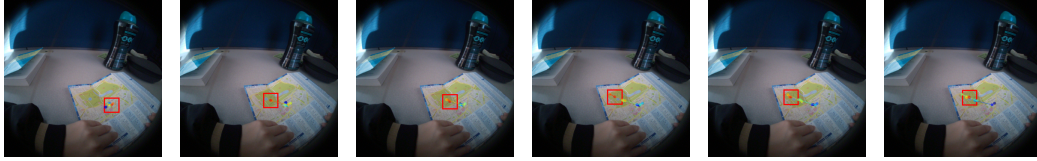


Figure 22: **All Modality Failure case: Searching on Map.** Failure case across 6 frames where all modalities and their combinations fail. Here the participant is searching for a particular name on map.

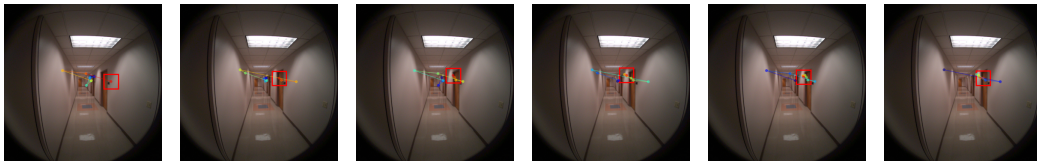


Figure 23: **All Modality Failure case: Reading room numbers while walking.** Failure case across 6 frames where all modalities and their combinations fail. Note that here the subject is asked to read room numbers while walking. The gaze pattern scans across the corridor before reading the room number. Note that the model was not trained on this kind of scenario.

E Comparison with Other Methods

In this section, we compare our method with other plausible approaches for this task, without considering power and compute constraints for always-on smart glasses. We include a discussion of alternative strategies, although we note that these methods are suboptimal both in terms of practical deployment and in how they handle multimodal inputs.

- **VLM-based methods:** This approach involves feeding each video frame into a vision-language model (VLM), with a large language model (LLM) backbone. Gaze can be incorporated in several ways such as by overlaying it onto RGB. However, this method is computationally intensive and currently impractical for on-device inference due to limited power and compute resources. Even when the VLM is not run on the device, streaming videos to a server still consumes significant power due to wireless communication cost.
- **Action recognition methods:** Reading has traditionally been addressed within the broader domain of action recognition, as discussed in the main paper. However, these models typically depend on sequences of full RGB and optical flow frames, whereas our approach uses a small crop of a single image patch processed by a much smaller model. This reliance on full-frame sequences results in significantly higher power consumption due to the increased sensing cost.
- **Alternative model architectures with gaze-only inputs:** We experiment with a few other alternative model architectures using gaze inputs alone, such as RNNs, CNNs, MLPs. While lightweight and suitable for on-device execution, these methods work with only gaze. This lack of sensor fusion poses challenges for detecting reading short texts on signs and objects or reading text while moving past them.
- **Ours:** We design our method to be simple such that the sensor fusion is seamless, and the computational burden is minimal. It employs low-power sensors, including eye tracking, IMU, and foveated image patches (low-power RGB). Notably, the RGB sensor dominates the sensory cost, while eye tracking and IMU are relatively inexpensive. The lightweight nature of our model allows it to run on-device with minimal computational requirements, while eliminating wireless communication costs. While further engineering could enhance performance, we believe this approach offers a robust and straightforward baseline to highlight the contributions of different sensors.

It is important to note that our contribution is not the model architecture itself, which builds up on standard transformer encoders, but in the identification and effective combination of different sensor modalities as inputs. The proposed architecture serves as a minimal, baseline framework for sensor fusion, with the emphasis on simplicity and ease of integration rather than architectural novelty.

While these are **not practical solutions**, we include the quantitative experimental results in the following subsections for completeness. For these experiments, we use our Reading in the Wild - Columbus subset, labeled as **RiTW Columbus** for measuring the zero-shot performance. In addition, we also evaluate on the EGTEA [4] dataset for the models with limited zero-shot capabilities. We use the dataset’s first training and test split. It is important to note that, **unlike our dataset, the EGTEA dataset only contains examples of reading long paragraphs from cooking recipes and lacks instances of reading short texts or hard negatives**, where text is present but not being read. This makes it easier for the existing models, as using RGB data alone is sufficient for most cases.

E.1 VLM based methods (Gaze + RGB)

To incorporate gaze into the input image, we simply overlay the gaze scanpath over the RGB image. We experiment using these settings.

- **VLM Model:** meta-llama/Llama-3.2-11B-Vision-Instruct [3]
- **Experiment:** Given an image with 2s eye gaze trajectory overlaid on top, prompt llm to determine if reading or not. We fed images with the full image resolution of 640 x 480 to the VLM.
- **Prompt:** This image shows eye gaze trajectory represented by connected green lines and a red circle, indicating where a person was looking when the image was captured. Determine if the person is reading or not reading. Answer only with 'reading' or 'not reading'.

Method	VLM [3]	Action recognition [2]	Alternative arch. with gaze only	Ours
1. Enabled modalities				
Gaze	✗	✗	✓	✓
RGB	✓	✓	✗	✓
IMU	✗	✗	✗	✓
Fusion	✓	✗	✗	✓
2. On-device Feasibility				
Number of parameters	11B	25M	1k	130k
Sensing cost (power)	high	high	low	low
RGB requirements (dominates sensing cost)	full RGB	full RGB video	-	foveated patch (5° FoV) (optional)
Real-time	✗	✗	✓	✓
Inference time (ms)	567.410	895.511 (incl. flow)	0.310	0.545
3. Performance				
Zero-shot capability	✓	✗	✗	✓
Acc / F1 on RiTW Columbus	76.7 / 65.6	-	-	82.9 / 88.8
Acc / F1 on EGTEA dataset	89.6 / 61.5	88.8 / 65.8	85.8 / 62.8	89.6 / 70.6

Table 7: **Comparison of alternative methods.** This table compares approaches for reading recognition, including (i) vision-language models (VLMs), (ii) action recognition models, and (iii) alternative architectures such as RNNs. Although all methods achieve reasonable performance on EGTEA (last row), these results are likely overestimated due to limited dataset diversity and the absence of hard negatives, leading to poor zero-shot generalization. Power and compute constraints of always-on smart glasses are not considered here—many of these models are impractical for on-device use. The high inference time of the action recognition model [2] is partly due to its reliance on optical flow computation.

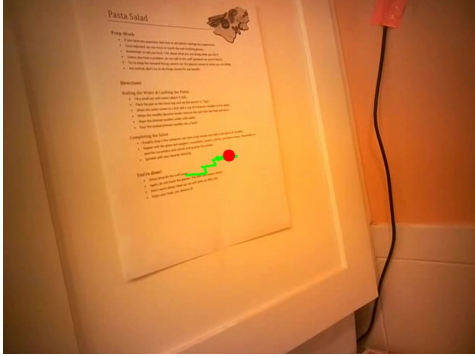


Figure 24: Reading

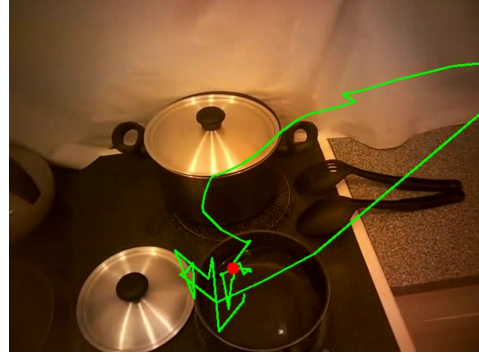


Figure 25: Not Reading

Figure 26: Example image inputs to the VLM (Llama-3.2-11B-Vision-Instruct [3])

Our method outperforms the VLM approach on both the Reading in the Wild Columbus subset and the EGTEA dataset. While the results are decent, we note that this is an impractical solution in terms of being able to run on-device at all times.

We further experiment with the settings to match the settings used in our model (e.g. RGB crop and gaze only) on the EGTEA dataset.

- RGB crop and gaze overlay on blank image: We use the same crop size and overlay the gaze onto a blank image to mimic our Gaze + RGB setting. The results are: F1 Score: 0.2648, Accuracy: 0.4622
- Gaze overlay on blank image: We overlay the gaze pattern onto an empty image to mimic gaze-only settings. The results are: F1 Score: 0.2480, Accuracy: 0.2564.

Evidently, the VLM has not seen such settings before and are unable to generalize zero-shot onto such prompts, which is not unexpected.

E.2 Action recognition based methods (RGB only)

To establish a baseline using action recognition method, we use the base two-stream I3D network [2]. The network has 25M parameters, is pretrained on Kinetics and fine-tuned on EGTEA (training set). We use both RGB and optical flow streams at full frames as input. Similar to before, we treat the class "Inspect/read recipe" as positive and the rest as negative.

The results are similar on the EGTEA dataset, but we again note that this is not a practical solution towards our use case which is real-time, on-device processing given the input and compute requirements. Action recognition models are large (both in terms of input requirements and parameters), and computing optical flow adds to latency. We further note that there are some works that try to incorporate gaze into action recognition models [5], but this further adds compute and goes further off tangent to our practical goal.

E.3 Alternative model architectures with gaze-only inputs (Gaze only)

Lastly, we also experiment with a few alternative model architectures with gaze inputs only, which we train on the EGTEA training split, including:

- **RNN.** First, we treat this as time series classification, and use a single layer RNN by treating gaze as time series. We use a single-layer GRU (32 channel dimensions, 1k parameters) followed by a linear layer to perform binary classification and train it with cross-entropy loss. We optimize using Adam with learning rate 0.003. To handle class imbalance during training we subsample the negative examples so that the training set is roughly uniform. The results are reasonable, as shown in Table 7.
- **1D CNN.** We also experiment with using a 1D CNN (3 layers of convolutions each of 32 channel dimensions, followed by a linear layer; 11k params total). This setup is similar to our model with gaze only input, but without the transformer layers. We find it to yield similar results (85.7 / 65.0 on EGTEA).
- **MLP.** Finally, we experimented with a 4-layer MLP with fully connected layers. This model resulted in overfitting.

We note that, while these methods may work on-device, the usefulness of gaze-only methods can be rather limited in cases where other modalities are more useful, such as reading short texts on signs and objects, or reading a text while passing by. In these cases, these methods are unable to naturally incorporate other sensors (RGB, IMU).

F Discussion

F.1 Reading medium classification using RGB on Seattle Subset

We also present the results for reading medium classification using the Gaze+IMU+RGB (10° FoV) model, as a pseudo upper bound. Notice that other mediums hardly get misclassified as digital, but there is still some confusion due to (i) small model (ii) gaze inaccuracy.

	(1)	(2)	(3)	(4)
(1) No read	0.88	0.03	0.02	0.07
(2) Print	0.04	0.74	0.01	0.21
(3) Digital	0.05	0.11	0.73	0.11
(4) Objects	0.05	0.32	0.01	0.62

Table 8: **Reading medium classification** using RGB.

F.2 Limitations and Future Work

Generalization and personalization. While our aim is to build a generalizable model that can capture the way people read, it does not take away the fact that different people read in different ways. Most prominently, we notice the reading speed varies greatly with language and the reader’s fluency, and the model sometimes fails when the reading becomes too slow.

For example, the average gaze velocity in the case of Chinese script is about two-thirds that of English. We did some simple personalization augmentations by simply scaling all velocity values to match the English case, and notice some improvements. This is a promising first step towards not only building reading detection models but also one that can be optimized towards a user.

Exploring eye tracking sensors. We have identified three modalities (gaze, RGB, and head pose), though we think that with eye tracking sensors gaze is not the only available information. Particularly, we also know that (i) the pupils dilate more (ii) people blink less when they are reading, and incorporating these cues may be interesting in the future.

Deployment and efficiency. We have shown that our model performs best by using all three combined modalities, but the model using gaze or RGB alone is also able to perform to good accuracy. In practical scenarios in wearable devices, power constraints sometimes restrict the use of all three modalities at all times. An interesting question here would be to know when to turn on/off each modality to maximize overall efficiency while still maintaining performance.

Gaze and human perception. Reading is one of the subset of activities where the eye gaze pattern clearly associates with the activity. More generally, it will be interesting to expand the questions we asked for reading towards more general human perception and ask: what does it mean to ‘look’ at something?

Imagine a scenario in which a person looks at a painting. The gaze can tell whether or not a person is looking at the painting, but is the person just glancing over it, or are they inspecting it to great detail? What aspect of the painting is the person looking at (color, shape, texture)? Are they examining critically or in awe, or just staring into blank space in introspection? These are questions that simply projecting where the eye is looking at onto the image does not yield a satisfactory solution, and will be interesting to investigate in the future.

References

- [1] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisper: Time-accurate speech transcription of long-form audio. *INTERSPEECH*, 2023.
- [2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.
- [3] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

- [4] 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.
- [5] Kyle Min and Jason J Corso. Integrating human gaze into attention for egocentric activity recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1069–1078, 2021.
- [6] 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.