# A Multimodal Approach to Investigate the Neural Mechanisms of Real World Social Vision

Arish Alreja[1,2,3,4], Michael J. Ward[4,5], Qianli Ma[6], R. Mark Richardson[4,7], Brian Russ[8], Stephan Bickel[9], Nelleke Van Wouwe[10], Jorge A. González-Martínez[4], Lisa Parker[11], Joseph Neimat[10], Charles Schroeder[8,12], L.P Morency[6], Avniel Singh Ghuman[1,4,13]

[1] Center for the Neural Basis of Cognition, Carnegie Mellon University and University of Pittsburgh
[2] Neuroscience Institute, Carnegie Mellon University
[3] Machine Learning Department, Carnegie Mellon University
[4] Department of Neurological Surgery, University of Pittsburgh
[5] David Geffen School of Medicine, University of California Los Angeles
[6] Language Technologies Institute, Carnegie Mellon University
[7] Department of Neurosurgery, Harvard Medical School and Massachusetts General Hospital
[8] Nathan Kline Institute
[9] Departments of Neurological Surgery and Neurology, Zucker School of Medicine at Hofstra University/Northwell
[10] Department of Neurological Surgery, University of Louisville
[11] School of Public Health, University of Pittsburgh
[12] Departments of Neurosurgery and Psychiatry, Columbia University
[13] Departments of Psychology, Neurobiology and, Psychiatry University of Pittsburgh

## Abstract

Eye tracking and other behavioral measurements collected from patient-participants in their hospital rooms afford a unique opportunity to study immersive natural behavior for basic and clinical-translational research, and also requires addressing important logistical, technical, and ethical challenges. Hospital rooms provide the opportunity to richly capture both clinically relevant and ordinary natural behavior. As clinical settings, they add the potential to study the relationship between behavior and physiology by collecting physiological data synchronized to behavioral measures from participants. Combining eye-tracking, other behavioral measures, and physiological measurements enables clinical-translational research into understanding the participants' disorders and clinician-patient interactions, as well as basic research into natural, real world behavior as participants eat, read, converse with friends and family, etc. Here we describe a paradigm in individuals undergoing surgical treatment for epilepsy who spend 1-2 weeks in the hospital with electrodes implanted in their brain to determine the source of their seizures.

This provides the unique opportunity to record behavior using eye tracking glasses customized to address clinically-related ergonomic concerns, synchronized direct neural recordings, use computer vision to assist with video annotation, and apply multivariate machine learning analyses to multimodal data encompassing hours of natural behavior. We discuss the acquisition, quality control, annotation, and analysis pipelines to study the neural basis of real world social and affective perception during natural conversations with friends and family in participants with epilepsy. We also discuss clinical, logistical, and ethical and privacy considerations that must be addressed to acquire high quality multimodal data in this setting.

## Introduction

Real world environments offer the opportunity to study behavior, and its physiological correlates, in ecologically valid settings. Technological advances in recent decades have enabled us to capture a greater number of behavioral and physiological variables in real time, with greater fidelity than ever before (Jacob Rodrigues et al., 2020). Doing so in real world environments allows us to study the relationship between physiology and natural behavior (Johnson and Andrews, 1996; Wilhelm et al., 2006).

Inpatient hospital environments are a unique real world setting to study the relationship between behavior and physiology. They feature monitoring of physiological data (EKGs, EMGs, heart rate, blood pressure, neural recordings, etc.) as part of standard care that can be augmented with behavioral monitoring. They also offer an opportunity to observe the relationship between behavior and physiology before, during, and after clinical events relevant to the patients' pathology. A deeper grasp of the relationship between behavior and physiology that accompanies clinical events has broad implications for diagnostics and our understanding of physiology-behavioral relationships in clinical disorders (Clark et al., 2019; Vigier et al., 2021; Wolf and Ueda, 2021). In addition, the hospital setting provides the opportunity to capture key

doctor patient interactions, whose salience for patients in such an environment cannot be overstated (Jhalani et al., 2005; Pickering et al., 2002). Modeling these interactions has deep implications in terms of understanding joint clinical decision-making, clinical information transfer, patient outcomes, and patient satisfaction in ways that cannot be replicated in controlled lab environments (Finset and Mjaaland, 2009; Girard et al., 2021; Kiesler and Auerbach, 2006; Muszynski et al., 2020; Weilenmann et al., 2018). Inpatient hospital environments also offer a compelling opportunity to advance basic knowledge by immersively studying natural behavior, such as interactions with friends and family, clinicians, eating, reading etc., in patients being monitored physiologically (Hogan and Baucom, 2016).

Real world environments encompass a multitude of physiological and behavioral processes unfolding at different timescales, which are affected by 'change events' in the environment itself(Shiffman et al., 2008). This makes them challenging to study. Successfully studying the relationship between behavior and physiology in such settings requires extracting meaningful insights from data that are rich, complex and heterogeneous in nature and in time. Inpatient hospital settings are subject to these considerations, as well as the additional complexity of hospital environments where unpredictable and potentially adverse events may unfold for patient-participants. In addition, they give rise to additional considerations that include patient privacy, well being, and the confidentiality of clinical information and doctor-patient interactions(Roter and Hall, 1989).

In this study, we present a methodology to collect behavioral and physiological data in epilepsy patients who are implanted with intracranial EEG electrodes and stay in the Epilepsy Monitoring Unit for 1–2 weeks for clinical identification of seizure focii and functional mapping. This clinical setting presents a unique opportunity to capture behavioral data (eye tracking using eye tracking glasses, audio, and video recordings) synchronized with neural activity recorded by intracranial EEG electrodes implanted in the patient's

brain, during real world social interactions with friends, family, clinicians and researchers. We discuss the privacy and ethical considerations that arise in this paradigm and how they can be addressed, as well as logistical challenges like fitting seizure prone patients, who have significant head bandaging protecting their implantation sites, with eye tracking glasses to collect data in a safe and robust manner. Finally, we describe data preprocessing and data fusion pipelines that can be used to construct a high quality multi-modal data set that blends real world social behavior and neural activity, allowing us to study the neural correlates of real world social and affective perception in the human brain.

# Materials & Methods

## Informed Consent

The informed consent protocols were developed in conjunction with a bioethicist (Dr. Lisa Parker) and approved by the Institutional Review Board of the University of Pittsburgh. Written informed consent was obtained from all participants. Audio and Video of personal interactions were recorded during experimental sessions. Although there are no reasonable expectations of privacy in public spaces like hospitals, our protocol incorporated several measures to ensure privacy considerations and concerns could be addressed based on the preferences of individual patients. Firstly, the timing of recording sessions was chosen based on clinical condition and patient preference, to ensure that they were comfortable with recording of their interactions with the visitors present (and/or expected to be present). Secondly, all visitors present in the room was notified about the nature of the experiment at the beginning of each recording session and given the opportunity to not be present if they wished. Thirdly, a notification was posted at the entrance of the patient room informing any entrants that an experiment where they might be recorded was underway and they could avoid going in if they chose to. Finally, at the end of each experimental recording, patients were polled to confirm their consent with the recording being used for

research purposes, and offered the option to have specific parts (e.g. a personal conversation) or the entire recording deleted if they wished. Thus, explicit permission for using recordings were acquired at the beginning and end of each session as this provides both affirmation of willingness to participate and provides the opportunity for participants to consider what was recorded before giving consent. Note that none of our participants thus far have asked to have recordings partially or fully deleted after the recording session was complete.

## Data Acquisition

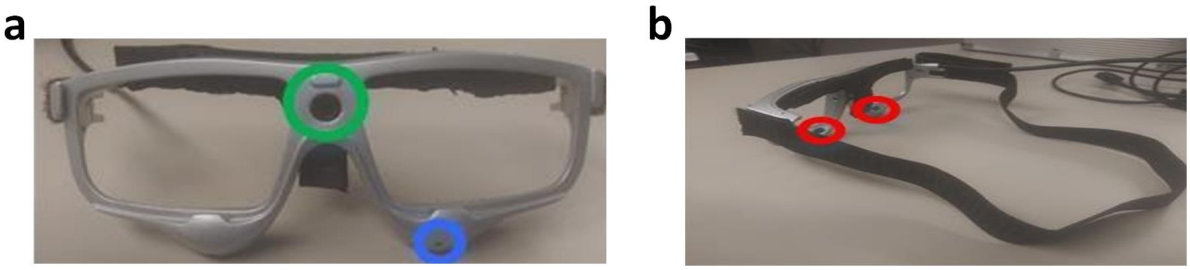


Figure 1: **a)** Front view of the SMI iView Eye Tracking Glasses with the green circle identifying the egocentric video camera which records the visual scene in front of the subject and the blue circle identifying the internal microphone which records audio. **b)** Back view of the SMI iView Eye Tracking Glasses showing the custom ergonomic velcro strap used to secure the glasses over the patient's bandaging, with red circles identifying two inward facing eye-tracking cameras.

Multimodal behavioral data (audio, egocentric video, and eye-tracking) as well as neural activity from up to 128 intracranial EEG contacts were recorded simultaneously during real world free viewing sessions in which patients wore eye-tracking glasses while they interacted with friends and family visiting them, clinicians and hospital staff responsible for their care, as well as researchers (Fig. 1). In addition, patients also engaged in other activities like eating meals, reading, and watching television. The type and duration of activities varied across different recording sessions.

Behavioral data were captured by fitting each patient with SMI's iView Eye Tracking Glasses (Fig. 1). An

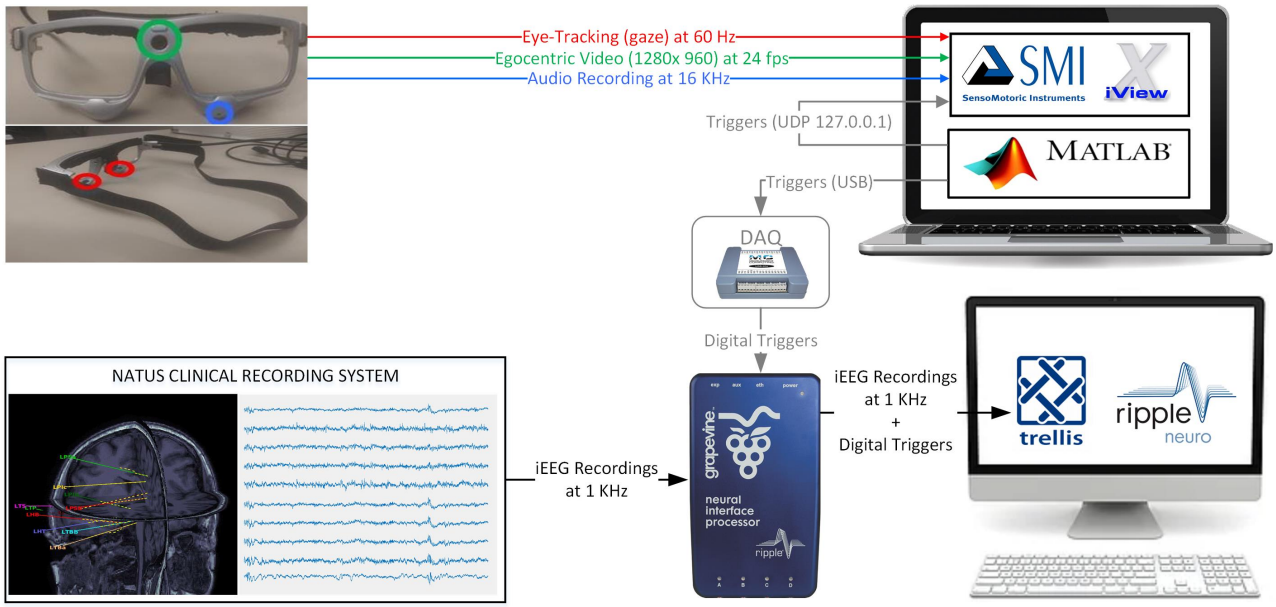


Figure 2: System level view of hardware setup to collect synchronized behavioral (top left) and physiological data (bottom left) from a patient during real world social interactions. The red, green and blue lines represent eye-tracking (60 Hz), egocentric video (1280x960 pixels, 24 Hz) and audio (16 KHz) recordings that encompass behavioral monitoring. The black lines represent intracranial EEG activity (1 KHz) from up to a 128 electrodes (visualized in MRI; with neural activity illustrated in adjacent plot), recorded using the Ripple Inc. Grapevine Neural Signal Processor. Digital Triggers, inserted in the eye-tracking using UDP packets over a sub ms latency local loopback connection and in the neural data stream using a DAQ, are shown in grey.

outward facing egocentric camera recorded video of the scene viewed by patients at a resolution of 1280 x 960 pixels at 24 frames per second (Fig. 1.a). Two inward facing eye tracking cameras recorded eye position at 60 Hz (Fig. 1.b). SMI's iView ETG server application, running on a laptop receives and stored streaming data for all three modalities from the eye-tracking glasses. Audio was recorded at 16 KHz (256 Kbps) using a microphone embedded in the glasses. This software also served as an interface for researchers to calibrate the eye-tracking glasses to each patient with a 3 point calibration procedure that enabled the accurate mapping of eye-tracking data to specific 'gaze' locations on video frames, and to initiate and stop the recording of behavioral data.

Electrophysiological activity (Field Potentials) was recorded from up to 128 intracranial EEG (iEEG) electrodes at a sampling rate of 1 KHz using a Grapevine Neural Signal Processor developed by Ripple Inc.

(Fig. 2). Common reference and ground electrodes were placed subdurally at a location distant from any recording electrodes, with contacts oriented toward the dura.

A MATLAB script, running on the same laptop as the SMI iView ETG Server software, broadcast numbered triggers every 10 seconds, injecting them simultaneously into the neural data stream via a DAQ connected to the Grapevine Neural Signal Processor's digital port and into eye-tracking event stream via SMI's iView ETG server application via a sub millisecond latency local loop back network connection using UDP packets (Fig. 2). These triggers were used to align and fuse the heterogeneously sampled data streams after the experiment, during the *Data Fusion* stage (see below for details).

**Best Practices for Behavioral Recording**

In each recording session, the capture of neural activity was initiated first, followed by eye-tracking glasses. Once the recording of all modalities was underway, the MATLAB script to generate and transmit triggers to all modalities was run. At the end of each recording session, the tear down sequence followed the reverse order, the MATLAB script was terminated first, marking the end of the recording, followed by shutting down the eye-tracking glasses, and the neural data recording was stopped last. Excess data prior to the first numbered trigger and after the last numbered trigger were discarded for all modalities.

Researchers instruct patients to avoid touching or nudging the glasses during a recording session to avoid disrupting the eye-tracking calibration completed at the beginning of a recording session. Such disruption can introduce systematic error(s) in eye gaze data captured after the disruption(s), although they can be mitigated with gaze correction (see *Data Preprocessing* for details). Despite such instruction, physical movement of eye-tracking glasses is possible if the patient inadvertently touches or moves the eye-tracking glasses due to discomfort or fatigue during a recording session. The potential for such an event

increases with the duration of a recording session. To minimize the risk of such error(s), we try to take a pause after an hour of recording. This means shutting down the recording, and starting a new one after a break of a few minutes. The interlude serves two purposes. First, it gives the patient a break from wearing the eye-tracking glasses, alleviating fatigue and discomfort. Second, initiating a new recording allows researchers to re-secure and re-calibrate the eye-tracking glasses, renewing the accurate mapping of gaze to the egocentric video. Although we prefer ≈1 hour recordings as a best practice, it is not always possible to maintain this practice depending upon the patient's preference or if they have a large number of visitors. In some cases, recording sessions may be longer.

## Ergonomic Modifications to Eye Tracking Glasses

Bandaging around the head protects operative sites where intracranial EEG electrodes have been secured with bolts from infection and secures the wires carrying electrical activity recorded by implanted intracranial EEG electrodes to clinical and/or research recording systems like the Grapevine Neural Signal Processor. The bandaging typically covered the patient's ears and rendered the ear-pieces on the eye-tracking glasses unusable. To overcome this challenge, we substituted ear-pieces with an elastic band and velcro patch. We attached velcro patches to the frame of the eye-tracking glasses and used an elastic band with velcro patches sown at each end to securely place the glasses over the patient's bandages, exercising care to avoid putting pressure around the operative sites (Fig 1.c). This ergonomic solution enabled correct and robust placement of eye-tracking glasses for each patient with flexibility to adjust to individual bandaging configurations. As an added measure to minimize the possibility of movement for eye-tracking glasses after during recording sessions, the USB cable from the eye-tracking glasses leading to the laptop was secured to the patient's clothing near their shoulder with a large safety pin to prevent the weight of the remaining length of the cable from pulling on and displacing the glasses during a recording session. Sufficient slack was left in the cable segment between the glasses and the fixation point on

the patient's clothing to allow for free head movement to prevent the secured cable segment pulling on (and potentially displacing) the eye tracking glasses.

## Data Preprocessing

The behavioral (eye-tracking, video, audio) and physiological (neural) data streams captured during a real world vision recording were preprocessed as follows before *Data Fusion* was initiated.

### Eye-Tracking

The eye-tracking data stream is composed of time series data sampled at 60 Hz, where each sample (referred to as an eye-tracking trace) contains a recording timestamp, an eye gaze location and is labeled by the SMI platform as belonging to a fixation, a saccade or a blink. Consecutive eye-tracking traces with the same label (fixation, saccade, or blink) are interpreted as belonging to a single eye-tracking 'event' of that type, whose duration is the difference in recording timestamps of the last and first eye-tracking traces in the block of consecutive traces with the same label (fixation, saccade or blink).

As an example, a set of 60 eye-tracking traces (amounting to 1 second of recorded activity), where the first 30 are labeled as fixation, the next 12 labeled as saccade, followed by the final 18 labeled as fixation would be interpreted as a fixation event ≈500 ms long (30 samples at 60 Hz), followed by a saccade event ≈200 ms long (12 samples at 60 Hz) followed by a fixation event ≈300 ms (18 samples at 60 Hz).

We developed custom python scripts that parse eye-tracking traces and construct logs of eye-tracking events for each recording session. In addition to the duration of each eye-tracking event, the median gaze location (median is used for robustness to outliers) was logged for each fixation event and the start/end gaze locations were captured for each saccade event. Blink traces are denoted by a loss of eye-tracking

(i.e. absence of gaze location) and as a result only the duration of blink events was tracked in the consolidated eye-tracking event logs.

Preprocessing of eye-tracking data also incorporates the detection and correction of systematic errors in eye gaze that can be induced by the movement of eye-tracking glasses during recording sessions (e.g. if a patient inadvertently touches and moves the glasses due to fatigue), which disrupts the calibration of eye-tracking glasses (see *Data Acquisition* for details). We screened for (and detected) such issues by manually viewing all experimental recordings using SMI's BeGaze application, which renders eye-gaze, audio and egocentric video together. Manual viewing of a behavioral recording makes it easy to spot systematic calibration issues induced by movement. For example, if the patient moves the glasses by inadvertently touching them, future fixations on the face of the person they are talking to will be altered in a consistent and visible manner. The SMI BeGaze application allows researchers to apply a manual correction to eye gaze (with a point and click procedure) at any time point in a recording, correcting all eye gaze data that follows. We review all recordings, apply manual corrections if needed and review the corrected sections to verify that the manually applied eye gaze corrections are indeed appropriate. Corrections to eye-tracking data precede preprocessing in such cases.

**Video**

Recordings of egocentric videos offer a broad range of visual stimuli, including objects, people and faces. Since the video recordings come from a camera mounted on the same glasses as the eye tracker it gives an egocentric view: the recorded videos are in the same direction as the participant's head and it rotates as the participant's head rotates. As a first research goal, we are interested to know what objects are present in the recorded videos. Of particular interests were faces and bodies, given the objective of examining social interactions. We processed videos to identify face regions and also body regions of peo-

ple recorded in the videos. As a secondary task, we were also interested to identify other non-face and non-body objects. Finally, for all face regions, we were interested to extract some higher-level measures about human visual behaviors, including head pose (including orientation and position of the head), eye gaze and facial expressions.

To automatically identify faces, people, and other objects, we used the automatic visual detection algorithm (Redmon et al., 2016). This software automatically processes each video frame and identify bounding boxes for each detected object (including labels for faces and people). A total of 1,450,462 video frames were processed this way. While there has been great progress in computer vision over the last decade, especially for automatic object detection, we decided to implement a second stage of annotation based on human judgements, to confirm the quality of the detection and correct any misdetection. Since manual annotations of all video frames is very time consuming, we annotated a single frame during each fixation given that fixations are defined by the lack of eye movement and thus patients look at the same thing during a fixation. Using the results of the data fusion (described below) between video frames and eye gaze information, we were able to a frame of each fixation using the common timestamps. We then use the FFmpeg (Tomar, 2006) software to extract the frames in the video according to the fixation timestamps resulting in a total of 125,836 fixation frames. The human annotator provides the coordinates of the bounding box of each face, object, or person in every frame. There can exist multiple faces and persons in a single frame.

As a final stage, we used the OpenFace software (Baltrusaitis et al., 2018), a facial behavior analysis toolkit using computer vision technology, to extract more higher-level information for the face regions. OpenFace gives for each face region information about (1) the position of 64 facial landmarks including eye, nose and mouth positions, (2) head orientation and position, (3) eye gaze direction and (4) facial expression

information encoded following the Facial Action Coding System standard (Friesen and Ekman, 1978).

**Audio**

Audio recordings from a microphone embedded in the eye-tracking glasses capture sound from the patient's perspective. The clarity of recorded audio is influenced by the loudness of sounds and the distance of the source from the patient. Since our objective involves examining social interactions, speech detection and speaker identification are behavioral markers of interest.

To detect time segments with speech in the audio recording and to diarize the audio (i.e. who spoke when?) we use state of the art deep learning speech detection (Lavechin et al., 2020) and speaker identification (Yin et al., 2018) pipelines available as part of an open source toolbox (Bredin et al., 2020). Even these state of the art models feature error rates which would be unacceptably high (for diarization in particular) for the use of their annotations as labels in analysis of neural data. In order to overcome this hurdle, we configure these models to be highly sensitive (leading to higher false positives, but very few false negatives) and then manually review model predicted time segments for speech and speaker identification, to identify and correct false positives. Outside of parameters that control the sensitivity of the deep learning models, the efficacy of speech detection and diarization is influenced by the loudness of the speakers themselves, as well as their distance from the patient (i.e, the microphone). This means that the patient's speech is the most reliably detected, while the quality of speech detection (and as a consequence, speaker identification) for other speakers may vary. As a result, we chose to collapse audio diarization into two categories, the patient and speakers other than the patient.

**Intracranial EEG recordings**

Response potentials and broadband activity were extracted from the raw iEEG recordings for statistical analysis using MATLAB™. Response potentials were extracted using a fourth order Butterworth band-pass ([0.2 Hz, 115 Hz]) filter to remove slow linear drift and high-frequency noise, followed by line noise removal using a fourth order Butterworth bandstop ([55 Hz, 65 Hz]) filter.

Broadband signal extraction involved two steps. First, the raw signal was filtered using a fourth order Butterworth bandpass ([1 Hz, 200 Hz]) filter followed by line noise removal using notch filters at 60, 120 and 180 Hz to obtain local field potentials. Next, Power spectrum density (PSD) between 70–150 Hz was calculated for the local field potentials with a bin size of 2 Hz and a time-step size of 10 ms using Hann tapers. For each electrode, the average PSD across the entire recording was used to estimate a baseline mean and variance of the PSD for each frequency bin. The PSD was then z-scored using these baseline measurements for each frequency bin at each electrode. Finally, the broadband signal is estimated by averaging the z-scored PSD across all frequency bins (excluding the line noise frequency bin at 120 Hz).

Artifact rejection heuristics were implemented during to avoid potential distortion of statistical analyses due to outliers. iEEG recordings were subjected to several criteria for inclusion in the study. Any recordings with ictal (seizure) events are not included in the study. Next, we evaluate the filtered iEEG data against three criteria that are applied to each sample i.e. each timepoint in iEEG recordings, which corresponds to 1ms of neural activity. These criteria were applied to the filtered iEEG signal for each electrode, as well as the averaged (across all electrodes) iEEG signal. The first criteria labels a sample as 'bad' if it exceeds 350 $\mu$V in amplitude. The second criteria labels a sample as bad if the maximum amplitude exceeds 5 standard deviations above/below the mean. The third criteria labels a sample as bad if consecutive samples (1 ms

apart at a 1000 Hz sampling rate) change by 25 $\mu$V or more. For the averaged iEEG signal, any sample satisfying any of these three rejection criteria is labeled as bad. Additionally, if more than 10 electrodes (out of a typical 128) satisfy the bad sample criteria for a particular sample, it is labeled as a bad sample. Less than 10% of the samples in experimental recordings were labeled as bad samples.

## Data Fusion

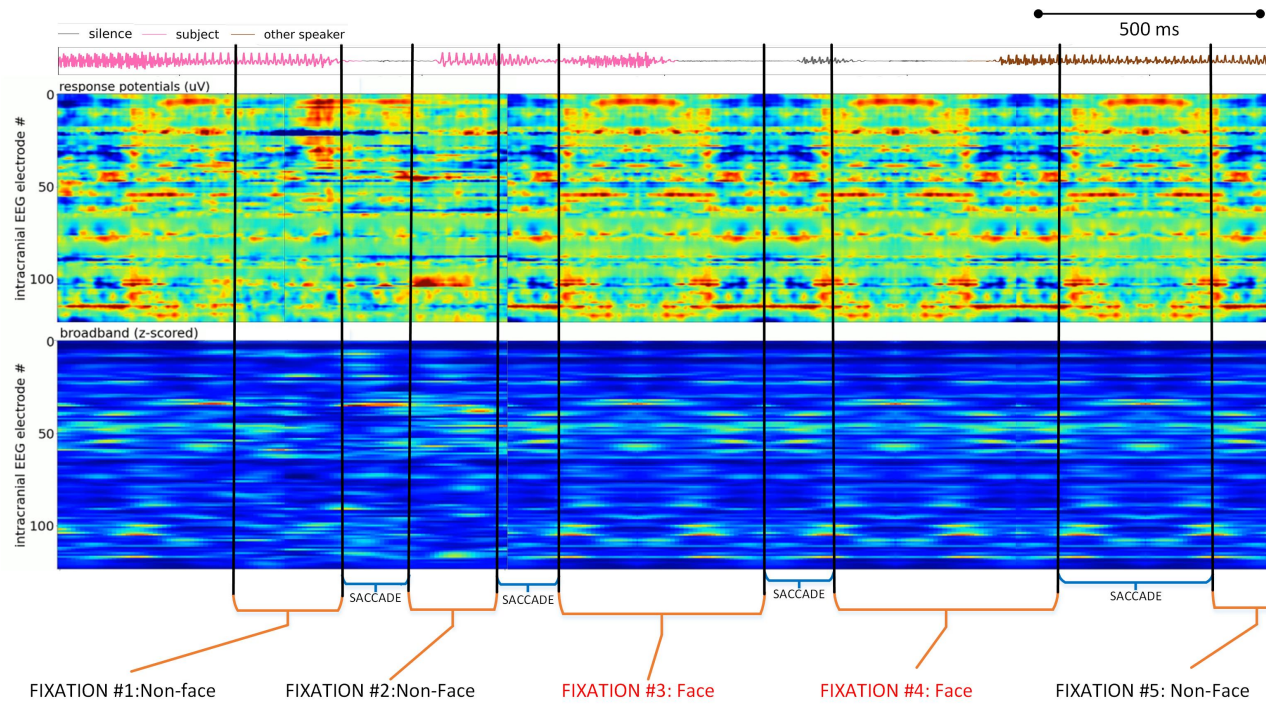


Figure 3: **Fused multimodal data set from a real world vision recording:** The audio waveform is shown on top, with colors denoting annotation labels that include silence, subject speech and speech from other speakers. Neural activity including response potentials and broadband high frequency activity for a 124 electrode montage are shown below the annotated audio. Vertical black line over the time series data (audio, neural activity) and colored time intervals below the neural activity demarcate different eye-tracking events (fixations and saccades).

Precise fusion of heterogeneous behavioral (eye-tracking, egocentric video and audio) and physiological (neural) data streams is essential for constructing a multimodal data set to answer our questions about the neural correlates of real world vision. In our methodology, eye-tracking serves as the reference modality against which behavioral (video, audio) and physiological (neural activity) data streams are aligned in time (Fig 3). Each eye-tracking event is mapped to a corresponding egocentric video frame. For fixation

events, we combine eye gaze location with bounding box locations/sizes from annotations for the egocentric video frame to determine what object (face or non-face) the patient is fixating upon. Each eye-tracking event is mapped to an auditory time segment and labeled as belong to a speech or silence segment, with additional labeling for speaker identity in the case of a speech segment. Finally, neural recordings are also aligned in time to eye-tracking events based on the temporal offset of eye tracking events and neural data, from trigger events which are injected in both data streams at 10 second intervals during recording sessions.

The quality of multimodal data sets assembled by the data fusion process described above is reliant on the quality of the heterogeneously sampled behavioral (eye-tracking, video, audio) and physiological (neural) data streams fed into the data fusion process. Acquisitional variability, if present and left undetected, can severely degrade the quality of fused data sets by introducing alignment issues and it is notable that dropped video frames and/or recording offsets are common. Our methodology includes cross-verification procedures that guard against such issues with careful examination of the raw data streams for each modality. These procedures assume that the raw data captured for any modality contains accurate and sufficient timing information to diagnose and correct such issues. As long as hardware/software systems in use meet this basic assumption about raw data capture, the cross-verification approach we describe should scale. Below, we detail two specific issues related that arise in our recordings using SMI Eye Tracking Glasses and illustrate how we addressed them to ensure data quality in the fused dataset.

**Sampling rate variability**

Variability in sampling rates is observed in engineered systems and can arise due to a variety of reasons ranging from temperature dependent variation in the frequencies of crystal oscillators that drive digital clock signals to propagation delays in circuit boards and circuitry running at different clock rates. If a fixed

sampling rate is assumed then these variations can accumulate as sampling drift over time and potentially lead to significant timing offsets over long periods of operation. These phenomena are addressed in engineered systems in various ways including using clocks far faster than the sampling rates desired and frequent resetting/calibration to minimize drift accumulation.

Here, we describe our approach toward detecting and removing such issues from the final multimodal data set that results from our data fusion procedures. We evaluated variability in the sampling rate of eye-tracking traces based on their timestamps. Since audio, video and neural data are anchored to eye-tracking events, minor sampling variability for eye tracking does not introduce any error as long as other data streams can be aligned to eye-tracking correctly. We evaluated the timing and mapping of all other modalities (audio, egocentric video and neural data) against eye-tracking. Specifically, we found the need to address sampling rate variability that arose in the egocentric video stream, so it could be reliably mapped to eye-tracking data.

The inter-frame interval for the video stream can vary systematically by small amounts from the rated 41.66 ms (for 24 fps) for a recording session. These deviations can be a source of error in the mapping of eye-tracking traces to video frames unless they are addressed during data fusion. A critical marker of this problem is an inconsistency between the number of frames present in the video and the number of video frames estimated from eye-tracking traces using Eq 1. It is important to note that this variability is not always accounted for in manufacturer software or documentation. The solution to this issue is relatively simple because the eye-tracking traces include a 'Video Time' column which has millisecond resolution. Instead of assuming a fixed frame rate as Eq 1 does

$$\underbrace{\text{Video Frame Number}}_{\text{in .avi file}} = \underbrace{\text{Video Time in seconds}}_{\text{from eye-tracking traces}} \times 24 \text{ frames per second} \tag{1}$$

We estimated video frame numbers corresponding to each eye-tracking trace using the 'Video Time' in them as follows

**Algorithm 1** Sampling rate variability resistant mapping of video frames to eye-tracking traces

```
Initialize trace_counter = 0, frame_num = 0, video_time = 0, eye_tracking_traces (from file)
while trace_counter<N do
    frame_num += round[(eye_tracking_traces[trace_counter].video_time - video_time) × 24]
    video_time = eye_tracking_traces[trace_counter].video_time
    eye_tracking_traces[trace_counter].video_frame = frame_num
    trace_counter = trace_counter + 1
```

**Addressing data gaps or corruption from behavioral modalities**

Loss or corruption of data during media recordings on embedded devices is a real world phenomena and a potential source of error for a fused multimodal data set that relies on precise alignment of multiple heterogeneously sampled modalities. As a result, our data fusion process pays close attention to identifying and characterizing such issues and addressing them to ensure data integrity. Here, we qualitatively describe different classes of issues observed in our data and how we address them to ensure data quality.

We observe missing frames in the video stream. Specifically, after correcting for sampling rate variability, we observed residual discrepancies between the number of frames that were expected and the number of frames present in the video files from experiments. By evaluating timestamps for each frame in the '.avi' files using low level video processing tools, we found that the lost frames were at the beginning of the video stream (i.e., the first $K$ frames of an $N$ frame video are missing) frames. We confirm this diagnosis with an additional form of verification, which used low level audio and video processing tools to manually blend audio and video streams with and without a correction for missing frames and visually verifying the absence of lip-audio synchronization issues in the resulting video. Finally, we obtained an additional point of manual verification by visualizing the ostensibly lost frames (decoders discard frames they deem corrupt when parsing a file, but they are present in the files) from the video file on a frame

by frame basis, confirming that they are corrupted/garbled. The specific pattern of error (first $K$ frames missing) observed with our experimental equipment (SMI iView Eye Tracking Glasses) may not replicate with other hardware, though given engineering constraints, other errors may arise instead. As an example, other eye-tracking glasses may have frame loss/corruption intermittently during a recording instead of at the beginning. However, our observations suggest that such issues may exist with other eye-tracking devices and data fusion pipelines should incorporate verification stages that can identify and correct such issues, with a preference for multiple modes of verification that are consistent with each other.

Blinks are a natural part of visual behavior and the eye-tracking traces denote them as such. Since eye-tracking is lost during blinks, there is usually no information about gaze, pupil dilation etc. available for blink events. We see blinks interspersed among fixations and saccades and they are typically a few hundred milliseconds long. However, we observed longer periods lasting several seconds in multiple recordings. To understand the phenomena better, we viewed the videos for periods where this happened, with gaze information overlaid using SMI's BeGaze software. We found these anomalous blinks to be masking a real phenomena where the patient may be looking out the corner of their eye, which takes their eye-gaze outside of the field of vision of the egocentric camera or upon occasion, take their eye-balls outside of the field of vision of the eye-tracking camera. Since the system cannot accurately capture visual behavior as it relates to the video in these conditions, it labels those periods as blinks. These scenarios are easy to spot during manual inspection because the eye-gaze location before and after the blink tends to be near the periphery of the video frame. These conditions are not a significant challenge for data quality, because they can be easily dropped from analysis. However, awareness of their existence is meaningful for data fusion pipelines.

## Subjects

A total of 6 human subjects (4 males, 2 females) underwent surgical placement of subdural electrocorticographic electrodes or stereoelectroencephalography electrodes as standard of care for seizure onset zone localization. Together electrocorticography and stereoelectroencephalography are referred to here as iEEG. The ages of the subjects ranged from 22 to 64 years old (mean = 37 years, SD = 13.47 years). No ictal events were observed during experimental sessions. All subjects gave informed consent in accordance with the University of Pittsburgh Institutional Review Board.

## Electrode Localization

Coregistration of grid electrodes and electrode strips was adapted from the method of Hermes et al. (2010). Electrode contacts were segmented from high-resolution postoperative CT scans of patients coregistered with anatomical MRI scans before neurosurgery and electrode implantation. The Hermes method accounts for shifts in electrode location due to the deformation of the cortex by utilizing reconstructions of the cortical surface with FreeSurfer™ software and co-registering these reconstructions with a high-resolution postoperative CT scan. All electrodes were localized with Brainstorm software (Tadel et al., 2011) using postoperative MRI coregistered with preoperative MRI images.

# Results

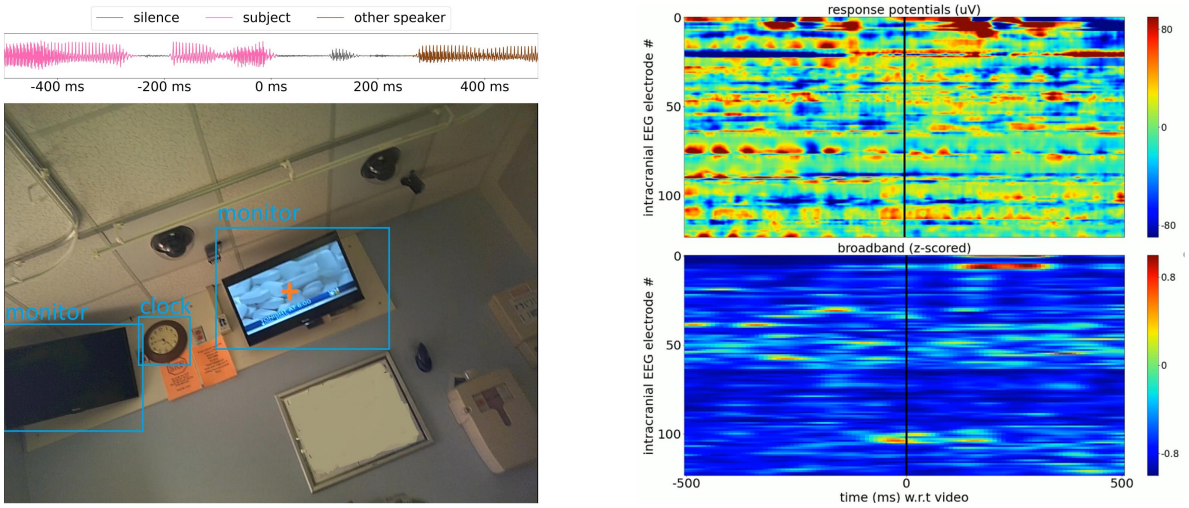


Figure 4: **Fused Multimodal (Audio, Egocentric Video, Eye Tracking and iEEG) Dataset:** This figure visualizes all modalities of the dataset together. On the left, an annotated audio snippet (top) and video frame (bottom) visualizes the world through the subject's eyes and ears as they interact with friends and family visiting them during a recording session. Speech/silence and speaker diarization labels color the audio signal on top and the video frame below is annotated with the orange '+' representing the subject's gaze reported by eye-tracking data and colored bounding boxes identify objects detected by computer vision models. The panel on the right visualizes preprocessed neural activity, which includes response potentials (top) and broadband high frequency activity (bottom) corresponding to the video frame across all intracranial electrodes implanted in this subject's brain.

We collected intracranial EEG recordings from patients in the Epilepsy Monitoring Unit (EMU) who wore SMI Eye Tracking Glasses as they went about their day interacting with friends and family visiting them as well as members of the clinical team. We used computer vision models to identify objects, faces and persons (bodies) in videos of the visual scenes in front of the patients during these sessions. Similarly, we used speech processing models to identify speech intervals and diarize the audio recorded from the internal microphone in the SMI eye tracking glasses. All annotation from computer vision and speech processing models was validated and corrected, if necessary, by human annotators to ensure data quality. Here, we show that fused multimodal datasets (see Fig 4 for a snapshot; see *Supplemental Video* for a dynamic version) which include annotated audio, eye-tracking, annotated video and intracranial EEG can be developed using this process. Such datasets can help advance visual neuroscience research beyond traditional experimental paradigms and explore the neural correlates of real world social vision.

## Behavioral Data

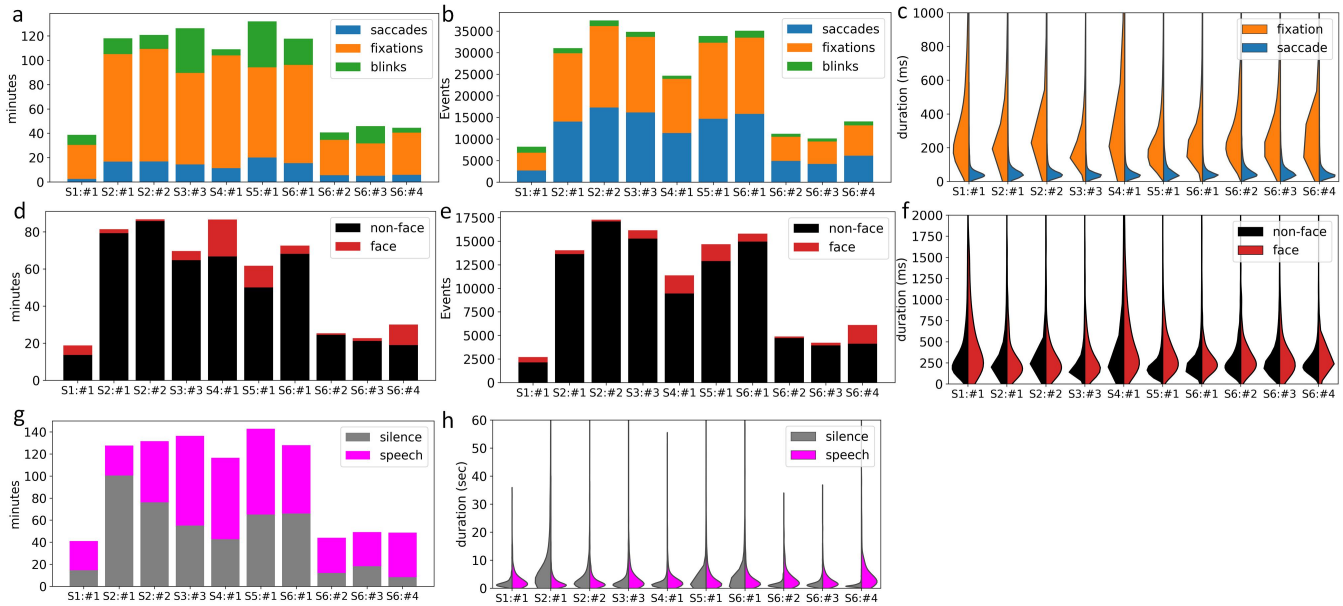


Figure 5: **Summary of dataset spanning visual behavior and the auditory environment: a)** The duration of each recording session (with multiple sessions for each subject) broken down by time spent in different visual behaviors (saccades, fixations and blinks). **b)** Similar to a), but counting distinct events for each visual behavior instead of time. **c)** Saccade and fixation duration distributions for each recording session. **d)** The amount of time fixations were on faces and non-face objects for each recording session. **e)** The number of fixations on faces and non-face objects for each recording session. **f)** Fixation duration distributions for face and non-face targets for each recording session. **g)** The duration each recording session broken down by time spent in silence and speech. **h)** Speech and silence segment distributions for each recording session.

We collected data from 6 patients across 11 different free viewing recording sessions which ranged from 41 – 143 minutes long and added up to a total of 16 hours and 48 minutes. Social contexts differed across recording sessions and sometimes within a recording session, in terms of the number of individuals present, the number of interactions they had with the patient and the nature of those interactions.

### Visual Behavior

SMI Eye Tracking glasses captured visual behavior, categorizing each moment as belonging to a saccade, fixation, or blink. Visual behavior varied depending upon the social context during recording sessions. Saccades usually accounted for 10–15% of the recording duration (Fig 5.a), even though they account for

nearly half the events (Fig 5.b) as a result of the saccade–fixation–saccade structure of the active sensing cycle, a contrast highlighted by the skew in the distribution of saccade durations and fixation durations (Fig 5.c). Saccades and fixations are not perfectly balanced due to the presence of blinks, which reflect loss of eye-tracking due to blinks and other reasons (e.g. noisy conditions, patients closing their eyes for brief periods or looking out of the corner of their eye during the recording sessions).

We identify fixation targets by combining gaze location from eye-tracking traces with bounding boxes from the video frame corresponding to each fixation. We categorize fixations as face and non-face fixations, reflecting our focus on the social aspects of real world vision. The social context during a recording session has a natural influence on the distribution of fixation targets. We found that patients fixated on faces less than 30–40% of the total time spent fixating during a recording session (Fig 5.d) even in the most social situations (e.g. EMU room full of multiple family and friends, with active conversations) time. The number of face fixations were an even smaller fraction of the total fixations (Fig 5.e), reflecting a skew in the distribution of fixation durations for the two fixation categories, where face fixations tend to be longer (Fig 5.f). Together, these observations indicate that even during the most social situations with familiar people, we look at the faces of people around us infrequently but when we do look at them we tend to hold our gaze a little longer.

**Auditory Context**

The SMI Eye Tracking glasses also recorded audio using an in-built microphone. We used deep learning models (Bredin et al., 2020) to do auditory scene analysis, augmenting it with manual annotation to ensure high reliability. Once again, depending upon the social context during each recording session, we observed varying levels of verbal discourse (Fig 5.g). We observed that speech could be detected from both the patient and others in the room, but the patient was reliably comprehensible due to their proxim-

ity to the microphone, whereas the volume and comprehensibility of the voices of other speakers would vary based on how close they were to the patient, making source identification more challenging even for manual annotation. To avoid potential confusion during manual annotation, we restricted speech diarization during supplemental manual annotation/verification to classifying speech as originating from the patient or other persons in the room. We find that more speech during a session (Fig 5.h) intervals is accompanied by a skew in the distribution of speech and silence intervals, which skews towards longer speech segments when there is more conversation.

**Behavioral Annotation: Reliability and its Cost**

Annotating video frames using automatic software is straightforward and fast, but it comes with a trade-off between speed and accuracy. The speed of the automatic software depends on the algorithms used for object detection. One of the most efficient algorithms is the YOLO algorithm (Redmon et al., 2016), which can perform object annotation (detection) at a rate of 45 fps on a NVIDIA K40 Graphics Processing Unit (GPU), or 5 fps on a standard CPU. This means annotating an hour of video should take only 32 minutes with a GPU, or close to three hours with a CPU.

The accuracy of automatic software is, however, significantly low. We measure the quality of the automatic annotation on one of the sessions (S4:#1) by comparing it with the human annotation. We find that the automatic software can only achieve a 10% Intersection over Union score (a measurement for evaluating object detection algorithms, higher the better, with a threshold of 100%). This means that the overlap ratio between the software's bounding boxes and the human annotators' bounding boxes is only 10%. Assuming the human annotator is reasonably more accurate, this shows that the automatic annotation software is rather inaccurate.

On the other hand, human annotation requires a larger amount of labor and time but produces annotations with significantly better quality. The human annotator annotated 125,836 frames out of 16 hours and 48 minutes of videos. The total time spent on annotating 125,836 frames is 103 hours. Thus, it takes an average of 6.13 hours for an experienced human annotator to annotate an hour of video.

For quality control, 15% of frames from the longest sessions (sessions from S2 and S6) were randomly sampled and verified by a second annotator. Then, an independent annotator randomly selected 1,000 frames from all sessions. Only five of the selected 1,000 frames have missing annotations, resulting in a 0.5% ratio of frames with missing annotations. Out of all people or faces that should be annotated, over 99.9% of them were annotated with a bounding box. Without the ground truth bounding box, the independent human annotator only found below twenty bounding boxes requiring re-annotating, less than 99% of all bounding boxes. This shows the significantly higher quality of human annotations over software annotations.

**Data Fusion Issues: Detected and Corrected**

Next, we show some results which motivate careful evaluation of the raw data for each modality before data fusion of heterogeneously sampled data streams from an experimental recording is attempted. Specifically, we describe and quantify alignment issues between eye-tracking and video data collected using SMI Eye Tracking glasses, that were identified and corrected (see *Methods for details)* during data fusion. We found two issues in the video stream, which would lead to misalignment between eye-tracking traces and the video frame they correspond to.

The first issue related to corrupted egocentric video frames at the beginning of each video recording, which were not recoverable. In data collected using SMI Eye Tracking glasses, the number of corrupted

frames varied for each recording, but we found that they were always at the beginning of the egocentric video recording. Crucially, the corrupted frames were discarded silently by video processing tools which returned the N+1$^{th}$ frame as the first frame (in the case of the first N frames being corrupted) and the issue was only root caused after careful examination of audio-video sync issues and timestamps embedded in the .avi file both verified its existence (see *Methods* for details). If left uncorrected, this issue would lead to incorrect mapping of eye-tracking traces to a video frame N+1 frames later than the egocentric video frame they corresponded to, which could lead to errors in annotation of fixations (e.g. as face or non-face fixation) across the entire video. The timing offset introduced due to this issue varied from 0 ms – 625 ms (Fig 6.a). After correction, the only impact of this issue is that eye-tracking traces/neural data for the first few frames that are corrupted and discarded cannot be used for analysis, which is a very minor loss.

The second issue related to variability in the average frame rate for egocentric video recorded from each session. We observed that for different sessions, the average frame rate of the recorded video was slightly above or below 24 frames per second. Eye-tracking traces are mapped to video frames using a 'Video Time' variable embedded in them. Estimating the video frame number corresponding to an eye-tracking trace using Eq. 1 which assumed a frame rate of 24 fps that was slightly higher or lower than the real frame rate of the video. The discrepancy led to an error between the estimated frame and the real frame corresponding to eye-tracking traces, which accumulated as the video recording progressed (Fig 6.b) and became visible with the eye-tracking traces mapping to far fewer/greater frames than were present in the video at the end of the recording. This issue was avoided by using the procedure defined in Algorithm 1, which was robust to these small variations in frame rate (see *Methods* for details). Both these problems co-occurred and addressing them as described in the *Methods* section gave us perfect consistency between the number of frames estimated in the eye-tracking traces and the number of frames present in the egocentric video. Lastly, we also evaluated audio and neural activity for similar alignment

inconsistencies with the eye-tracking logs and found no issues with alignment.

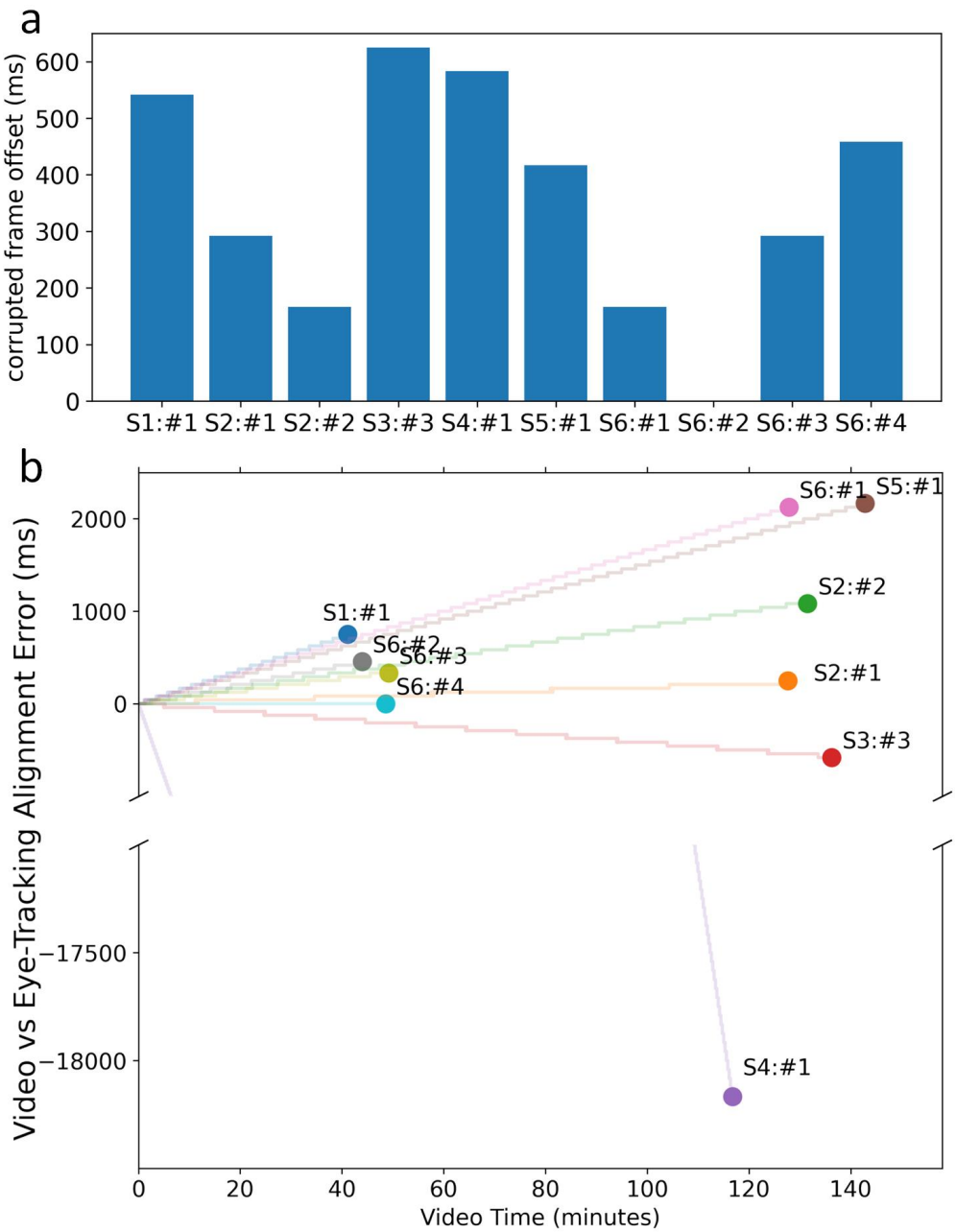


Figure 6: **Potential effects of video frame corruption and video frame rate variability on the accuracy of data fusion:** Timing error for each recording session introduced in the alignment of eye-tracking events and their corresponding video frames **a)** if corrupted frames at the beginning of each video file are not detected, quantified and accounted for in the alignment procedure. This is a fixed error that affects all eye-tracking events in a session. **b)** If the procedure to map frames to eye-tracking traces is not robust to variations in video frame rate/inter-frame intervals video frame offset. This is a time varying error which accumulates over the duration of a recording (scatter points indicate the final accumulated error at the end of each recording session) and its rate of accumulation (slope of shaded lines in the background) depends upon how big the deviation in video frame rate is from 24 fps. We observe deviations as small as 1 frame (41.67 ms) over a 49 minute recording for S6:#4 and as large as 432 frames (18 seconds) over a 2 hour recording for S4:#1.

## Neural Correlates of Real World Social Vision

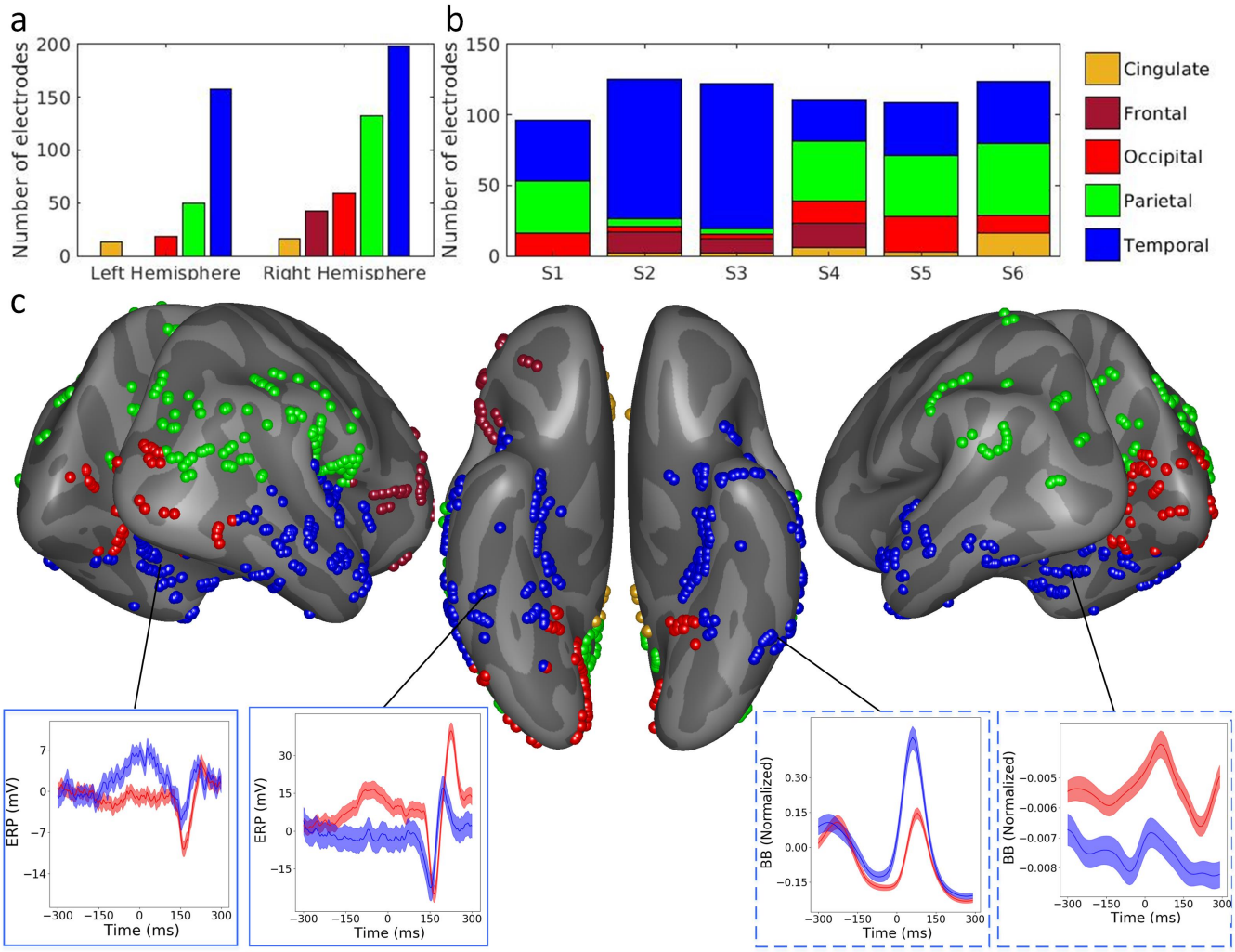


Figure 7: **a)** Cortical distribution of the 686 intracranial EEG electrodes from 6 subjects over different lobes across the Left and Right Hemispheres per the Desikan Killiany atlas (Desikan et al., 2006). **b)** Per subject electrode distribution across different cortical regions **c)** Visualization of locations of electrodes from all subjects on an inflated cortical surface with ventral, lateral (left and right), posterior and anterior views. Average fixation locked neural activity from electrodes sampled across all subjects and recording sessions. The colors of the boxes correspond to the lobe of the cortical location being sampled and the outlines denote the neural signal that is visualized (solid lines denote Fixation Response Potentials (uV) and dashed lines denote Fixation Response Broadband Activity (a.u.)). The average fixation locked response to face (red) and non-face objects (blue) is shown for each cortical location. One notable result is that differences in the neural response between face and non-face fixations appear prior to fixation onset, suggesting predictive activity/"pre-saccadic preview"(Buonocore et al., 2020; Huber-Huber and Melcher, 2021)

The number and cortical locations of intracranial EEG electrodes from which neural data were recorded varied by patient with a total of 686 cortical locations distributed across the temporal, parietal, occipital,

frontal and cingulate areas of patients (Fig 7.a, b).

Finally, we align neural activity recorded from intracranial EEG electrodes to the composite behavioral (eye-tracking + visual behavior + auditory context) log using digital triggers embedded in the neural and the eye-tracking data streams. This final step allows identification and extraction of neural activity corresponding to individual eye tracking events (saccades, fixations and blinks). Our analysis of real world vision is anchored to fixations and Fig 7.c visualizes average Fixation Response Potentials (FRPs) and Fixation Response Broadband (FRBB) for face and non-face fixations from several of the 686 intracranial EEG electrodes for which real world vision data were collected. Typical aspects of the response potential (e.g. enhanced N170 for faces, particularly in ventral temporal cortex locations) and response broadband signal are well represented for electrodes from multiple lobes suggesting the alignment of neural activity and eye-tracking events is robust and provides a key ”proof-of-principle” for this real world paradigm.

## Discussion

We have demonstrated the feasibility of combining neural recordings from intracranial EEG electrodes with eye-tracking, video and audio recordings collected using eye tracking glasses and annotated using computer vision and speech models to generate robustly fused multi-modal data sets from real world recording sessions in an inpatient hospital environment. Fusion of visual behavior with neurophysiological recordings enables investigation of the neural correlates of real world social vision and affective perception. The efficacy of our methodology is validated in the context of real world social vision by fixation locked neural activity for face and non-face fixations from ventral temporal electrodes, which shows category selective neural signatures that are also typical in traditional visual neuroscience experiments. The preliminary data also raise several potential opportunities for the enrichment of behavioral and phys-

iological data collection as well questions of significant interest for clinical and translational research.

## Enriching behavioral monitoring

### Higher fidelity capture of visual behavior

From analyzing the data sets presented here, three natural opportunities to improve the capture of visual behavior are apparent. The first entails higher fidelity data acquisition for behavioral data streams that we already capture. The eye-tracking glasses used in this study feature a single egocentric video camera operating at 24 frames per second with a resolution of 1280 x 960 pixels capturing a 60° (horizontal) by 46° (vertical) region of the field of vision, with 2 eye-tracking cameras operating at 60 Hz. Increasing the spatial resolution of the video camera in pixels, improving the temporal resolution of both eye-tracking and video and capture a larger fraction of the field of vision can aid in better tracking of visual behavior over a more complete portion of the field of the vision. The second opportunity requires adding a new data modality (head position) using an Inertial Measurement Unit (IMU), that can provide tracking for the physical frame of reference corresponding to each video frame. The third opportunity involves considering the addition of depth perception information for eye-gaze, which may potentially be supported by the addition of a second egocentric camera. A review of available research grade hardware (Cognolato et al., 2018) provides an account of the capabilities of several research grade devices, which can be evaluated for their suitability with respect to each of these possibilities.

### Aural scene capture

Analysis and annotation of the auditory scene recorded using the in-built microphone embedded in the eye-tracking glasses reveals the potential advantages of capturing the aural scene as well as the limitations of having a single microphone physically attached to the patient. Adding high definition microphones

in the room and recording the auditory scene completely offers the opportunity to go beyond speech and speaker detection and into speech recognition, and its conversion to text (Hannun et al., 2014; Ren et al., 2019; Zhang et al., 2018) allowing the use of language models that can add an additional behavior modality for semantic and sentiment analysis (Kiritchenko et al., 2014).

**From monitoring visual behavior to visual monitoring of behavior**

Heavily monitored inpatient hospital environments like an epilepsy monitoring unit are typically equipped with cameras that allow clinical care teams to monitor patient behavior. The same video streams also capture the physical behavior of other individuals (e.g., doctors, nurses, family) who are present. These video streams hold the potential to add two additional behavioral modalities to the multi-modal data set we have described. The first modality is affective behavior, for the patient and other individuals present, extracted using facial analysis tools like OpenFace (Baltrusaitis et al., 2018). The second modality is physical behavior using tools like OpenPose (Cao et al., 2019) and DeepLabCut (Insafutdinov et al., 2016; Lauer et al., 2021; Mathis et al., 2018, 2021; Nath* et al., 2019) which may enable us to explore the relationship between physiology and behavioral phenomena like interpersonal synchrony (Delaherche et al., 2012).

## Enriching physiological monitoring

Inpatient hospital environments feature the monitoring of a wide variety of physiological data like EKGs, EMGs, heart rate, blood pressure, neural recordings, saliva samples, urine samples as well as clinical events, as part of standard care. A richer physiological data set than the one presented here, that contains a greater number of the physiological modalities can combine powerfully with behavioral markers to allow pursuit of highly relevant clinical and translational research questions.

## Implications for Clinical and Translational Research

Real world social interactions in an inpatient hospital setting include Doctor-Patient interactions(Girard et al., 2021; Muszynski et al., 2020), which include interactions with neurosurgeons in the case of patients in the epilepsy monitoring unit. Capturing physiological and behavioral data corresponding to these interactions offers a unique opportunity to understand how clinical decision making in these dyadic interactions is affected by different circumstances based on factors like the severity of clinical issues involved, the presence of family, the patient's mental health etc. A deeper understanding of the relationship between patient physiology and behavior that accompanies clinically important interactions has profound implications for clinical practice, patient outcomes and patient satisfaction (Korsch et al., 1968).

## Ethical Considerations

Capturing real world behavior brings with it a number of privacy considerations that are compounded by a clinical environment (Berg et al., 2001; Parker, 2002). These concerns are partially mitigated by the fact that a hospital is an open environment, with hospital staff coming and going and rooms often left open to the hallway and, in many cases, the room being under video and audio monitoring. Furthermore, putting on, calibrating and wearing the eye tracking glasses is a physical reminder that behavior is being captured.

Nonetheless, in order to do ethical research, participant comfort and protection of privacy is critical. In conjunction with a bioethicist (Dr. Lisa Parker) we developed a multistep process to ensure that consent was clear and participants were comfortable with the data that was captured being used for research purposes. Specifically, first informed consent is acquired that explicitly describes the fact that video and audio data of natural behavior will be acquired. Second, during the setup, the participant is reminded about the fact that video and audio recordings will be acquired and guests/visitors in the room are also

made aware of this fact. Third, a note is placed on the door to alert anyone who enters the room that research video and audio recordings are taking place in the room. Finally, at the end of the session, the participant is asked to consider the events that happened during the recording and explicitly consent to them being used for research. In addition, separate consent/release is acquired for using these video and audio recordings in figures for presentations. In this paradigm, we sought to study natural real world social interactions and thus avoided recording doctor-patient interactions or clinical events. For studies that seek to understand doctor-patient or clinical events, these protections and privacy concerns become even more acute and participants should be reminded when acquiring both pre- and post- session consent that the video/audio recordings will include sensitive clinical information.

The question of data sharing for recordings that are inherently not de-identifiable is an additional issue to consider. Processed data (annotations with identifiable information removed, for example audio diarization and generic aspects of the computer vision annotations) could likely be shared openly as long as substantial care was taken to assure de-identification. Sharing raw data is a bigger challenge and would require additional layers of consent such as consent procedures used when creating public behavioral databases, though it even with this level of protection care must be taken given the potential sensitive nature of the recordings in a clinical environment. Thus, at most, well curated snippets of raw data may be publicly shareable, and sharing of raw data would likely have to be done under IRB approval at both institutions with a data use agreement.

## Conclusion

Our current methodology augments physiological monitoring with behavioral monitoring in experimental recording sessions in the epilepsy monitoring unit. Scaling behavioral monitoring to around the clock can

enable the collection of multi-modal clinical data sets that are large enough for cutting edge machine learning techniques like deep learning to be pressed into service to learn relationships between behavior and physiology, including potential behavioral and physiological markers that are predictive for clinically significant events like seizures, cardiac events, respiratory events and others.

## Funding

This study was supported by the National Institutes of Health under award R01MH107797 and R21EY030297 (to A.S.G.) and the National Science Foundation under award 1734907 (to A.S.G) and 1734868 (to L-P.M). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the National Science Foundation.

## Acknowledgements

We thank the patients for participating in the iEEG experiments and the UPMC Presbyterian epilepsy monitoring unit staff and administration for their assistance and cooperation with our research. We thank Nicole Silverling and Taylor Gautreaux for their help with annotations of the egocentric video and Katherine Lindsay for her help with audio annotations.

Conflict of Interest: None declared.

## References


Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE, 2018.

Jessica W Berg, Paul S Appelbaum, Charles W Lidz, and Lisa S Parker. *Informed consent: legal theory and clinical practice*. Oxford University Press, 2001.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. Pyannote.audio: Neural building blocks for speaker diarization. In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 2020.

Antimo Buonocore, Olaf Dimigen, and David Melcher. Post-saccadic face processing is modulated by pre-saccadic preview: Evidence from fixation-related potentials. *Journal of Neuroscience*, 40(11): 2305–2313, 2020.

Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

Rosie Clark, James Blundell, Matt J Dunn, Jonathan T Erichsen, Mario E Giardini, Irene Gottlob, Chris Harris, Helena Lee, Lee Mcilreavy, Andrew Olson, et al. The potential and value of objective eye tracking in the ophthalmology clinic. *Eye*, 33(8):1200–1202, 2019.

Matteo Cognolato, Manfredo Atzori, and Henning Müller. Head-mounted eye gaze tracking devices: An overview of modern devices and recent advances. *Journal of Rehabilitation and Assistive Technologies Engineering*, 5, 2018.

Emilie Delaherche, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux, and David Cohen. Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*, 3(3):349–365, 2012.

Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.

Arnstein Finset and Trond A Mjaaland. The medical consultation viewed as a value chain: a neurobehavioral approach to emotion regulation in doctor–patient interaction. *Patient education and counseling*, 74(3):323–330, 2009.

E Friesen and Paul Ekman. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 3(2):5, 1978.

Jeffrey M. Girard, Alexandria K. Vail, Einat Liebenthal, Katrina Brown, Can Misel Kilciksiz, Luciana Pennant, Elizabeth Liebson, Dost Öngür, Louis-Philippe Morency, and Justin T. Baker. Computational analysis of spoken language in acute psychosis and mania. *Schizophrenia Research*, 2021. ISSN 0920-9964. doi: https://doi.org/10.1016/j.schres.2021.06.040. URL https://www.sciencedirect.com/science/article/pii/S0920996421002528.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

Dora Hermes, Kai J Miller, Herke Jan Noordmans, Mariska J Vansteensel, and Nick F Ramsey. Automated

electrocorticographic electrode localization on individually rendered brain surfaces. *Journal of Neuroscience Methods*, 185(2):293–298, 2010.

Jasara N Hogan and Brian R Baucom. Behavioral, affective, and physiological monitoring. In *Computer-assisted and web-based innovations in psychology, special education, and health*, pages 3–31. Elsevier, 2016.

Christoph Huber-Huber and David Melcher. The behavioural preview effect with faces is susceptible to statistical regularities: Evidence for predictive processing across the saccade. *Scientific reports*, 11(1): 1–10, 2021.

Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, pages 34–50. Springer, 2016.

Mariana Jacob Rodrigues, Octavian Postolache, and Francisco Cercas. Physiological and behavior monitoring systems for smart healthcare environments: A review. *Sensors*, 20(8):2186, 2020.

Juhee Jhalani, Tanya Goyal, Lynn Clemow, Joseph E Schwartz, Thomas G Pickering, and William Gerin. Anxiety and outcome expectations predict the white-coat effect. *Blood pressure monitoring*, 10(6): 317–319, 2005.

P Johnson and DC Andrews. Remote continuous physiological monitoring in the home. *Journal of telemedicine and telecare*, 2(2):107–113, 1996.

Donald J Kiesler and Stephen M Auerbach. Optimal matches of patient preferences for information, decision-making and interpersonal behavior: evidence, models and interventions. *Patient education and counseling*, 61(3):319–341, 2006.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M Mohammad. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762, 2014.

Barbara M Korsch, Ethel K Gozzi, and Vida Francis. Gaps in doctor-patient communication: I. doctor-patient interaction and patient satisfaction. *Pediatrics*, 42(5):855–871, 1968.

Jessy Lauer, Mu Zhou, Shaokai Ye, William Menegas, Tanmay Nath, Mohammed Mostafizur Rahman, V. Di Santo, Daniel Soberanes, Guoping Feng, V. Murthy, G. Lauder, C. Dulac, M. Mathis, and Alexander Mathis. Multi-animal pose estimation and tracking with deeplabcut. *bioRxiv*, 2021. URL https://www.biorxiv.org/content/10.1101/2021.04.30.442096v1.

Marvin Lavechin, Marie-Philippe Gill, Ruben Bousbib, Hervé Bredin, and Leibny Paola Garcia-Perera. End-to-end Domain-Adversarial Voice Activity Detection. In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 2020.

Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie W. Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 2018. URL https://www.nature.com/articles/s41593-018-0209-y.

Alexander Mathis, Thomas Biasi, Steffen Schneider, Mert Yuksekgonul, Byron Rogers, Matthias Bethge, and Mackenzie W. Mathis. Pretraining boosts out-of-domain robustness for pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1859–1868, January 2021.

Michal Muszynski, Jamie Zelazny, Jeffrey M Girard, and Louis-Philippe Morency. Depression severity assessment for adolescents at high risk of mental disorders. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 70–78, 2020.

Tanmay Nath*, Alexander Mathis*, An Chi Chen, Amir Patel, Matthias Bethge, and Mackenzie W Mathis. Using deeplabcut for 3d markerless pose estimation across species and behaviors. *Nature Protocols*, 2019. URL https://doi.org/10.1038/s41596-019-0176-0.

Lisa S Parker. Information (al) matters: Bioethics and the boundaries of the public and the private. *Social Philosophy and Policy*, 19(2):83–112, 2002.

Thomas G Pickering, William Gerin, and Amy R Schwartz. What is the white-coat effect and how should it be measured? *Blood pressure monitoring*, 7(6):293–300, 2002.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

Yi Ren, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Almost unsupervised text to speech and automatic speech recognition. In *International Conference on Machine Learning*, pages 5410–5419. PMLR, 2019.

Debra L Roter and Judith A Hall. Studies of doctor-patient interaction. *Annual review of public health*, 10 (1):163–180, 1989.

Saul Shiffman, Arthur A Stone, and Michael R Hufford. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.*, 4:1–32, 2008.

François Tadel, Sylvain Baillet, John C Mosher, Dimitrios Pantazis, and Richard M Leahy. Brainstorm: a user-friendly application for MEG/EEG analysis. *Computational Intelligence and Neuroscience*, 2011, 2011.

Suramya Tomar. Converting video formats with ffmpeg. *Linux Journal*, 2006(146):10, 2006.

Marta Vigier, Katherine R Thorson, Elisabeth Andritsch, Herbert Stoeger, Leonie Suerth, Clemens Farkas, and Andreas R Schwerdtfeger. Physiological linkage during interactions between doctors and cancer patients. *Social Science & Medicine*, 284:114220, 2021.

Sonja Weilenmann, Ulrich Schnyder, Brian Parkinson, Claudio Corda, Roland Von Kaenel, and Monique C Pfaltz. Emotion transfer, emotion regulation, and empathy-related processes in physician-patient interactions and their association with physician well-being: a theoretical model. *Frontiers in psychiatry*, 9:389, 2018.

Frank H Wilhelm, Monique C Pfaltz, and Paul Grossman. Continuous electronic data capture of physiology, behavior and experience in real life: towards ecological momentary assessment of emotion. *Interacting with Computers*, 18(2):171–186, 2006.

Alexandra Wolf and Kazuo Ueda. Contribution of eye-tracking to study cognitive impairments among clinical populations. *Frontiers in Psychology*, 12:2080, 2021.

Ruiqing Yin, Hervé Bredin, and Claude Barras. Neural Speech Turn Segmentation and Affinity Propagation for Speaker Diarization. In *19th Annual Conference of the International Speech Communication Association, Interspeech 2018*, Hyderabad, India, September 2018.

Zixing Zhang, Jürgen Geiger, Jouni Pohjalainen, Amr El-Desoky Mousa, Wenyu Jin, and Björn Schuller. Deep learning for environmentally robust speech recognition: An overview of recent developments. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(5):1–28, 2018.