# FacialMotionID: Identifying Users of Mixed Reality Headsets using Abstract Facial Motion Representations

1<sup>st</sup> Adriano Castro KASTEL Security Research Labs Karlsruhe Institute of Technology Karlsruhe, Germany adriano.castro@student.kit.edu 2<sup>nd</sup> Simon Hanisch Centre for Tactile Internet (CeTI) Technical University Dresden Dresden, Germany simon.hanisch@tu-dresden.de

4<sup>th</sup> Thorsten Strufe KASTEL Security Research Labs Karlsruhe Institute of Technology Karlsruhe, Germany thorsten.strufe@kit.edu 3<sup>rd</sup> Matin Fallahi KASTEL Security Research Labs Karlsruhe Institute of Technology Karlsruhe, Germany matin.fallahi@kit.edu

Abstract—Facial motion capture in mixed reality headsets enables real-time avatar animation, allowing users to convey non-verbal cues during virtual interactions. However, as facial motion data constitutes a behavioral biometric, its use raises novel privacy concerns. With mixed reality systems becoming more immersive and widespread, understanding whether face motion data can lead to user identification or inference of sensitive attributes is increasingly important.

To address this, we conducted a study with 116 participants using three types of headsets across three sessions, collecting facial, eye, and head motion data during verbal and non-verbal tasks. The data used is not raw video, but rather, abstract representations that are used to animate digital avatars. Our analysis shows that individuals can be re-identified from this data with up to 98% balanced accuracy, are even identifiable across device types, and that emotional states can be inferred with up to 86% accuracy. These results underscore the potential privacy risks inherent in face motion tracking in mixed reality environments.

*Index Terms*—Privacy, Biometric Data, Facial Motion Data, Mixed Reality, Eye Gaze, Face

# 1. Introduction

*Mixed Reality* (MR) promises to fuse the real and digital worlds. This implies the universal tracking of MR users to create precise digital twins of them. Their appearance, voice, and motions are captured and streamed onto digital avatars. The newest generation of MR headsets (e.g., Apple Vision  $Pro^1$  and Meta Quest  $Pro^2$ ) already integrate face and eye tracking to animate the faces of these digital avatars (see Figure 1 for an example). Integrating facial and eye motions



Figure 1. A user wearing a mixed reality headset with facial motion tracking. Their avatar mimics their facial expressions.

improves social interactions in MR, as subtle non-verbal cues can now be transmitted to a dialogue partner. Currently, we are still in the early adopter stage of this technology as only a handful of applications such as VRChat<sup>3</sup> or virtual YouTubing (using a virtual character to create videos) make use of facial motions. Nonetheless, with the advancement of MR, facial motion tracking is expected to become a standard feature of future MR devices.

3. https://hello.vrchat.com/

<sup>1.</sup> https://www.apple.com/apple-vision-pro/

<sup>2.</sup> https://www.meta.com/de/en/quest/quest-pro/

However, sharing facial motion data in MR poses a potential privacy risk because facial motions are a behavioral biometric trait. It may yield both identity and attribute disclosure risks: An attacker could use the facial motion data from the avatar shared in MR to perform privacy inferences like identification or employ attribute inferences, like emotion recognition.

Imagine a user visiting a digital store in the Metaverse wearing an MR headset. The user has a generic avatar that does not reveal their identity, and the avatar's facial motion tracking is turned on by default. Without the user's knowledge, the store owner can collect their facial motion data by observing the avatar's facial animations. The store owner can use this data to identify the user, determine if they have visited the store before, and recognize their facial expressions to see which items they like. Thus, the user shares much more private information than they realize.

Although many behavioral biometric traits, such as gait [1], voice [2], and eye gaze [3], are already known to be privacy sensitive, this remains an open question for facial motions. Therefore, we seek to understand whether individuals can be identified from facial motion data and whether emotional states can be inferred. To this end, we designed and conducted a study in which we recorded 116 participants using three types of MR headsets. Each participant attended a maximum of three sessions, with each session being approximately a week apart. During each session, participants were recorded with two types of headsets while performing a set of verbal and non-verbal tasks, with multiple repetitions of each task.

We then performed privacy experiments using the collected dataset by investigating the general identification using three different biometric recognition models. Besides general identifiability, we investigate if participants can be re-identified across sessions and different MR headset types. Further, we look at emotion recognition and examine which tasks are best suited for identification.

The main contributions of this paper are as follows:

- We recorded a novel facial motion dataset, which for the first time allows the investigation of associated privacy risks.
- We demonstrate that the identification of individuals is possible from facial motion data alone.
- We show that re-identifying people across sessions and different MR headsets is possible.
- We confirm that emotion recognition from abstract facial motion data can be performed with high accuracy.

The paper is organized as follows. In Section 2, we explore the related work of facial motion identification, and then describe the background in Section 3. We then first describe the general study design in Section 4 before we describe our concrete study implementation in Section 5. Afterwards, we evaluate the study by performing identification experiments in Section 6, and subsequently discuss them in Section 7. We end the paper with a short conclusion in Section 8.

# 2. Related Work

In the following section, we will present the related research on identifying individuals in MR through analysis of facial motion. The primary focus of our research lies in the development and analysis of methodologies for the identification of facial motion from video data, the detection of facial expressions, the identification through eye gaze, and the identification of individuals from MR motion data.

### 2.1. Facial Motion Identification

Some preliminary studies have been conducted on the identification of individuals based on facial motion, with the majority of these studies focusing on video data.

Benedikt et al. [4] employed 3D videos of faces to assess the distinctiveness of facial motion for biometric authentication. The trajectory of these facial motions is then represented within the Eigenvector space of diverse facial expressions. Their findings indicate that non-verbal tasks may not be as effective in terms of identification from facial motions as verbal tasks. Zhang et al. [5] performed a similar study, in which they collected 3D videos of participants speaking a passcode 10 times. The system demonstrates an impressive capacity to identify the participant from the dynamic features of the video, achieving a 96% accuracy rate with 77 participants. Haamer et al. [6] collected a video dataset of 61 participants performing various emotion tasks. They then show that participants can be identified using the videos recorded.

Moreira et al. [7] utilized a neuromorphic sensor, an advanced device capable of capturing precise alterations in individual pixels, to record the facial expressions of 40 participants while reciting nursery rhymes. They can show that identification is possible with accuracies as high as 96%.

The existing literature suggests that the identification of individuals through facial motion is feasible for both facial expression and speaking tasks. However, given that the majority of studies employ video data, it remains uncertain whether identification can be achieved exclusively through the analysis of facial movements alone, since face recognition is possible on static face images. Additionally, the question remains open whether individuals can be identified across multiple sessions via facial motion data.

#### 2.2. Facial Expression Recognition

One field of study that has focused on facial motion analysis is facial expression recognition. The objective of facial expression recognition is to categorize the emotions displayed by the individual captured on video [8]. Zhao et al. [9] propose a lightweight model to extract the displayed emotion from face images. Wen et al. [10] use an attention network to perform emotion recognition and achieve stateof-the-art performance. Furthermore, Chen et al. [11] have employed the differences between a neutral face and an expressive face to enhance the learning of different face expressions. To improve generalization in their face recognition model, Zhang et al. [12] propose learning an identityindependent representation of facial expressions using deviation learning. This involves subtracting a person's identity, established by a face recognition model, from their facial expression embedding.

Lee et al. [13] investigate facial expression recognition using a face mask that measures facial deformation, rather than via videos.

Facial expression recognition has also already been investigated in the context of MR by Chen et al. [14] in a study in which extra cameras have been integrated into an existing MR headset. Additionally they used an external camera to capture the part of the face which is not hidden behind the MR headset. They then show that they can achieve a facial expression recognition accuracy of 95%.

Facial expression recognition shows that facial motion data is useful for more than just direct social interactions between people. However, this information should also be considered private, and individuals should have the choice of when and how they share their emotions.

## 2.3. Eye Gaze

Eye gaze was recognized as a privacy-sensitive topic some time ago and has also drawn attention as a possible behavioral biometric trait for authentication. Lohr et al. [15] showed that they could identify 269 subjects with a mean EER of 4.72% using the SBA-ST dataset [16], which was captured with a dedicated eye tracker. They further improved their method in EyeKnowYouToo [17], which is the current state-of-the-art model for user authentication based on eye gaze. They achieved an EER of 3.66% at a sampling rate of 1000 Hz and an EER of 8.77% at a sampling rate of 125 Hz. In a later study, Raju et al. [18] investigated the performance of eye gaze authentication on the GazeBasedVR [19] dataset and showed that short-time authentication works well but that the EER increases to 10% for longer sessions.

Shao et al. [20] aim to create an eye-gaze identification system in MR that is independent of the content shown to users. They use two encoders: one for content and one for eye gaze. They achieved an F-score of 92%. Asish et al. [21] use eye gaze features of 34 people performing four different tasks for identification in *Virtual Reality* (VR).

As the privacy-sensitive nature of eye gaze data has been recognized, the first studies [22]–[24] seeking to anonymize it have emerged. Common methods of anonymization include adding noise or smoothing the eye gaze trajectories.

Eye gaze is useful not only for authentication, but also for foveal rendering. Foveal rendering is a selective rendering process that increases the level of detail in the section of the image at which the user is looking. Several studies [25]– [27] attempt to predict eye gaze to enable foveal rendering.

The research on eye gaze data showcases the dual nature of behavioral biometric data, as both privacy inferences, as well as desired applications like authentication and foveal rendering are possible with it.



Figure 2. The data sharing pipeline of facial motion data captured by MR headsets.



Figure 3. A blendshape named "MouthRight" being activated on an MR avatar from 0 to 1 through interpolation.

#### 2.4. Mixed Reality Identification

In recent years, the subject of identifying people using motion data recorded by MR headsets has gained traction, and multiple studies have been published on the topic. Among the first of these studies, Miller et al. [28] recorded 511 participants watching 360-degree videos in VR. The researchers demonstrated a high identification rate of 95% using the head and controller motions. Liebers et al. [29] demonstrated that identifying individuals is possible by combining the head orientation and eye gaze of 12 people captured with a MR headset. Moore et al. [30] investigated which VR tasks are most effective for identification, once again using headset and controller motions. They found that identification success depends on the VR content used. Nair et al. [31] used a large-scale dataset of people playing Beat Saber<sup>4</sup> and demonstrated their ability to identify players in a pool of over 50,000 people with 94% accuracy using 100 seconds of headset and controller motion data.

# 3. Background

Here, we briefly describe the background for MR motion tracking and biometric recognition required for this work.

<sup>4.</sup> https://www.beatsaber.com/

### 3.1. Mixed Reality Tracking

Facial Motion Tracking: The MR headsets used in our study rely on camera-based face tracking. Inward-facing infrared cameras capture the eyes and mouth of the person wearing the headset. This video data is then transformed into a symbolic representation which is shared via applications on the MR headset. See Figure 2 for the full data sharing pipeline of facial motion data. For facial motions, the data is represented as blendshapes. Blendshapes are a type of interpolated animation, also known as morph target animation. In this type of animation, the neutral state and deformed version of an object are stored for each blendshape. Then, for each frame of the animation, the object's vertices are interpolated between the neutral and deformed versions. An example of a blendshape for an MR avatar is the right part of the mouth (see Figure 3). In the neutral state, the mouth is symmetrical; in the deformed state, it is pulled to the right side of the face. All intermediate states can be created via interpolation. The blendshapes defined by the MR headsets are usually based on the Facial Action Coding System (FACS) [32], [33]. The former is a system that defines and describes all distinguishable facial movements, so-called action units. These action units are derived from anatomy, and with them complete expressions can be recognized objectively. Two examples of such action units are "Cheek Raiser" and "Lip Corner Puller" that together can be interpreted as the expression of happiness [34].

**Eye Tracking:** The user's eye gaze is captured via infrared cameras positioned inside the MR headset. The video is then converted into gaze direction and eye position data.

**Motion Tracking:** For the motion tracking in MR headsets there exist two main approaches. Inside-out tracking describes the approach in which multiple cameras on the outside of the headset are used to establish its position. The second approach is light house tracking in which one or multiple static light houses emit sequences of infrared light which are registered by infrared sensors on the surface of the headsets and the controllers. The headset and controllers can then compute their distance and orientation in relation to the light houses. When comparing the two approaches, insideout tracking is less precise but easier to use than lighthouse tracking.

#### 3.2. Biometric Recognition

**Biometric traits** (sometimes **biometric characteristics** [35]) are properties of a human that either capture what a human is (e.g. face, iris) or how a human behaves (e.g. voice, gait, heartbeat). The former are known as biological, the latter as behavioral biometric traits.

**Biometric recognition** is the process of inferring the identity or specific attributes of an individual from its biometric data. For inferring the identity we consider two cases. **Authentication** entails the verification of a claimed identity given the input of one fresh observation and a template representing the class of that claimed identity. The main threat

to biometric authentication is impersonation: An adversary succeeding a verification attempt for another individual's identity. **Identification** of a given observation produces the most likely candidate (or: list of top-k candidates) from all learned classes that represent individuals, together with their respective classification confidence. Complementary to such identity disclosure, **Attribute inference** is a privacy threat in which a specific private attribute (e.g., age, sex, medical condition) of an individual is inferred from the biometric data.

### 4. Study Design

In this section, we describe the design of our study to investigate identity and attribute disclosures from abstract facial motion data. We first explain the general rationale before providing a more detailed explanation of the tasks used and the selected recording schedule.

#### 4.1. Design Rationale

The main goal of our study is to investigate whether identifying individuals from their facial motion data is possible. To allow biometric recognition systems to train on the data and recognize identifying patterns, we require a large number of samples. Therefore, we require numerous repetitions and task executions involving a diverse group of participants. Additionally, we aim to determine whether facial motion data is a stable biometric factor over time; therefore, we will record multiple sessions with each participant. Lastly, we want to investigate whether facial motion data generalizes well when different devices are used to capture facial expressions. Therefore, we record our participants using multiple device types that integrate facial motion tracking.

We see the main application of facial motion data for animating digital avatars as speaking to other people and displaying emotions. Consequently, we focus on tasks involving two types of categories, namely speech and emotional expression for data collection. As mentioned in Section 2.2, emotion recognition has been shown to work previously. Hence, we integrate it into the study to test collected data and to compare results. Since facial motion data will likely be used in combination with eye gaze and head motion data—and as these are readily available in the common MR headsets—we also collect these.

#### 4.2. Recording Procedure

We chose to record our participants over the course of three separate sessions, with each session being approximately a week apart from each other. In the first session, participants first answer a short questionnaire about demographics before the actual recording starts. During each session, we record each participant performing the same set of tasks with two different MR headset types. We chose to keep one headset type the same throughout all sessions,



the expression starting with a neutral face after pressing the button.

Figure 4. The participant performs Figure 5. The participant performs the expression starting with a neutral face after pressing the button.

whereas the respective other headset was alternated between the remaining two in each session. This allowed us to record all participants using three different headsets. Due to the change in the second headset, we split our participants into two groups, A and B, to keep track of which second headset had to be used in each session.

# 4.3. Tasks

We designed a task-based study in which participants performed predefined tasks sequentially. An overview can be seen in Table 1. At the beginning of the study, one tutorial task was performed for each task type. To cover the described applications, we selected verbal tasks, in which participants read a given text aloud, and non-verbal tasks, in which participants mimic a facial expression. Studies such as [4], [7] have demonstrated that verbal tasks contain the most identity cues in facial motion, unlike non-verbal tasks. Therefore, the predominant task category we selected is verbal tasks.

First, the participant is shown the current task. Then, the participant starts the actual recording phase for the task by pressing a button. During the recording phase, the participant performs the task. The recording phase is ended by pressing the same button again. All tasks and their repetitions are presented to the participant in a random order. There are four repetitions for each task in the first session and five repetitions for each task in the second and third sessions. The reduction of repetitions in the first session allows time for the questionnaire.

Non-verbal Tasks: We presented the non-verbal tasks using emoticons that displaying three different facial expressions: happiness, anger, and fear. See Figure 4+5 as an example for a non-verbal task. This abstract representation should encourage participants to perform the facial expressions as they normally would rather than closely mimicking the avatars shown to them. Therefore, we did not use highfidelity digital avatars. We instructed participants to mimic the non-verbal tasks shown to them by starting with a neural facial expression and to then transition into the shown facial expression. An animation of the emoticon changing from neutral to the target expression illustrates this process.



Figure 6. An example of a verbal task in which the participant is uttering the nursery rhyme "Sing a Song of Sixpence".

Verbal Tasks: During the verbal tasks (see Figure 6), participants are asked to utter words and sentences. Lu et al. [36] have shown that words and groups of sentences that contain a large number of phonemes are best suited for identification. A phoneme is the smallest unit of sound which makes a lexical difference in a language. Additionally, Moreira et al. [7] have already shown that reciting nursery rhymes are suitable for facial motion identification. Therefore, we selected nursery rhymes for the verbal tasks because they contain various repetitive phonemes. To select the nursery rhymes, we used a list<sup>5</sup> of common English nursery rhymes. To keep the verbal task short, we prepared the list by splitting all rhymes, such that each part is at most four lines long. Next, we counted the phonemes of each nursery rhymes and selected the top three with the highest count. Out of these selected nursery rhymes, we selected one word of each that contained the highest amount of phonemes, constituting the word tasks.

TABLE 1. OVERVIEW OF THE DIFFERENT TASKS WHICH THE PARTICIPANTS PERFORMED IN THE STUDY. V: VERBAL, NV: NON-VERBAL

ID	Туре	Task	Repetitions
0	v	sixpence (word)	4/5
1	v	dinosaurs (word)	4/5
2	v	muffin (word)	4/5
3	v	Sing a Song of Sixpence (rnhyme)	4/5
4	v	Dinosaurs (nrhyme)	4/5
5	v	The Muffin Man (nrhyme)	4/5
6	nv	happiness	4/5
7	nv	anger	4/5
8	nv	fear	4/5

# 5. Study Implementation

The study was conducted between January 22 and February 14, 2025. It took place in a dedicated laboratory that

<sup>5.</sup> https://www.bbc.co.uk/teach/school-radio/articles/z4ddgwx

contains multiple small booths specifically designed for user studies, and an office for their supervision. We divided each study day into 12 slots, with each day ranging from 8:30 am to 6:15 pm. Since we aimed for a study duration of approximately 30 minutes, an equal time allocation was assigned to each slot. To compensate for unexpected duration times, we added a 15-minute break between each slot. At each slot, two individuals participated simultaneously — one from group A and another from group B. As each booth contained a door, each participant could perform the study without any disturbances.

#### 5.1. Ethics

The data collection was approved by the ethics commission of the Karlsruhe Institute of Technology (research project "Privacy of Facial Motions") and was conducted in accordance with the Declaration of Helsinki. Participants were paid based on their time of participation at an hourly rate of  $14 \in$ . Additionally, participants received a flat bonus of  $2 \in$  or  $3 \in$  for participating in the second and third sessions, respectively. We obtained informed consent from all participants for the data collection and processing.

#### 5.2. Apparatus

During the study, we used four MR devices, namely two Meta Quest Pros, one Pico 4 Enterprise<sup>6</sup>, and one HTC Vive Pro Eye<sup>7</sup> with the Facial Tracker add-on<sup>8</sup>. All of these devices support eye and facial tracking in addition to standard head and controller tracking. Moreover, the devices and their tracking are supported by Unity, the Game Engine that we used to implement the application for our study. While the first device type is designed for both augmented and virtual reality, the other two are purely VR devices. Since we only require VR, the three types of devices were deemed suitable for our experiments.

The study was implemented as a Unity application since all selected MR devices supported it. Unity Engine v2021.3.32f1 was utilized for development, as it was the most recent long-term support version supported by all headsets and their tracking APIs. We created a scene for each device, as they required individually configured XR cameras and device specific code to activate their motion tracking.

To be able to access the motion data of the devices and store them, we utilized several Unity packages that allowed the interaction with the **APIs** of the devices. For the Meta Quest Pro we used the Meta Movement SDK v71.0.1 including the Meta XR Core v71.0.0 and the Meta XR Interaction SDKs v71.0.0<sup>9</sup>. For the Pico 4 Enterprise we used the PICO Unity Integration SDK v2.5.0<sup>10</sup>. And

9. https://developers.meta.com/horizon/documentation/unity/moveoverview/

10. https://developer.picoxr.com/document/unity/?v=2.5.0

for the HTC Vive Pro Eye we used the VIVE OpenXR Plugin v2.0.0<sup>11</sup> with addition of the VIVE SRanipalRuntime v1.3.1.1 and the OpenXR Plugin v1.9.1 for the facial tracker. Both our Meta Quest Pros used during the study had identical software and runtime as well as OS versions, namely v71.0.0 and SQ3A.220605.009.A1 respectively. The Pico 4 Enterprise ran on version v5.9.9, and the Vive's eye and lip camera versions were v2.41.0-942.e3e4 and v50100 in corresponding order.

#### 5.3. Recruitment

We recruited 116 participants (45 female, 71 male; age mean 23.6 years, std 4) with the help of the KD2Lab panel of the Karlsruhe Institute of Technology.The distance between two subsequence sessions was between 4-16 days (participants per session 1: 116, session 2: 83, session 3: 49). Of the participants, 67 were native German speakers, while the rest reported a different mother tongue. 67 describe themselves as ambiverts, 26 as extroverts, and the remaining 23 as introverts.

The participants were assigned to their respective group at random. While group A used the HTC Vive Pro Eye in addition to their assigned Meta Quest Pro in the first session, group B started with the Pico 4 Enterprise. In the second session, group A then received the Pico 4 Enterprise instead of the HTC Vive Pro Eye, and group B vice versa. In the third session, group A and B each returned to their first headsets. Thus, each participant who participated in all sessions used each device at least once and the Meta Quest Pro three times.

# 5.4. Session Procedure



Figure 7. A participant performing the tasks with the HTC Vive Pro Eye.

11. https://github.com/ViveSoftware/VIVE-OpenXR-Unity

<sup>6.</sup> https://www.picoxr.com/global/products/pico4e

<sup>7.</sup> https://www.vive.com/sea/product/vive-pro-eye/overview/

<sup>8.</sup> https://developer.vive.com/us/hardware/facial-tracker/

For the first session, our participants required more thorough guidance and support. We began by introducing the study and explaining the procedure, emphasizing the data collection process and its purpose. Then, we started a timer to keep track of their study duration, which was relevant for their payment at the end. Then, we assigned each participant a random pseudonym to be used for the remainder of the study.

Next, we escorted each participant to their assigned room. Each participant was given an information sheet with details about the study, a data protection agreement, and a survey. The survey collected information about the participants' age, sex, origin, self-assessed personality traits, English proficiency, and mother tongue. After completing the survey, the participants watched a short introductory video showing them how to use the MR headsets and their respective calibration procedures.

After watching the tutorial videos, the participants were brought to the booth with their first headset. We helped them become accustomed to the headset and to perform the eye calibration. Thereafter, the participants started the Unity application and, thus, performed the tasks shown through their MR headset. When they completed the tasks with the first headset, they were brought to the second one, where we repeated the procedure. At the end, the participants filled in a short online survey to receive their payment with their own payout token assigned through the experiment organization. To reduce any possible bias in the data due to headset order, the order of the headsets was inverted for each group of participants. See Figure 7 for an example how the participants performed the study.

In subsequent sessions, participants did not have to fill out the survey or data protection sheet again. Although we asked the participants if they wanted to watch the eye calibration tutorial videos again, they usually skipped them since they remembered how to perform the tasks. Additionally, the subjects usually skipped reading the study information sheet from the first session. They were usually brought directly to the headsets and performed the study as described above.

#### 5.5. Troubleshooting

During the study, there were some difficulties. For the first recording day (22.01.2025) we encountered a problem for the facial motion recording of the HTC Vive, and as a consequence the HTC Vive recordings for the first day contain less blendshapes then the following recordings. Another problem we encountered with the HTC Vive was that for some of the audio recordings the recording frequency was higher than configured, though this was unproblematic since our data processing approach presented in Section 5.6 is robust against it. The eye calibration of the Meta Quest Pro devices turned out to be challenging, as it would regularly finish unsuccessfully. This seemed to be more frequent with participants wearing glasses, yet it also happened with non-glasses wearers. In such problematic cases, we helped the participants adjust the lenses and the position of the headset

on their heads — it did help a relative number of cases, but not all of them. Due to these problems, the quality of the eye tracking for the Meta Quest Pro suffered. Another challenge was that both the eye and face tracking of the Meta Quest Pro devices tended to suddenly malfunction in between participants. This happened once per Meta Quest Pro device, and was unfortunately only discovered at the end of the day. Due to this issue, we lost 19 recordings.

### 5.6. Data Processing

Upon the completion of each participant's session, our Unity project generated a unique directory containing the relevant data and metadata gathered during it. This included the unsegmented face, eye, and head motion data, as well as the execution order and timestamp range of each task repetition, a microphone recording along with its metadata, and a log file.

Since the facial and eye motion data formats exported by the MR devices are not exactly the same, a unification step was necessary. See Table 2 for the exact mapping for each MR headset. For n-to-1 mappings from the devices to the unified format, we use the mean of the directions. One example of this is the CheekPuff blendshape. The HTC Vive and Meta Quest Pro support CheekPuff for both sides of the face, while the Pico 4 Enterprise only returns one CheekPuff blendshape.

As our study consisted of tasks, we partitioned the unsegmented data of each participant into individual tasklevel segments. Moreover, we segmented the aforementioned task-level segments which belonged to text tasks further into word- and phoneme-level segments.

**Task-Level Segmentation:** First, the data was segmented by task. To achieve this, we used the timestamp ranges stored during each task repetition. When a participant started a task, a timestamp was saved to mark the start of execution. Then, when the participant finished the task, a second timestamp was saved to mark the end. Since we stored the timestamp of when each sample of motion data was collected, we could identify which samples belonged to which task repetition in each motion data file.

#### **Text-Level Segmentation:**

We further segmented the verbal tasks into words (nursery rhymes only) and phonemes. To accomplish this, we aligned the speech recordings collected during task execution with the transcript of the performed task. We used a force alignment model to automatically perform this process on all verbal tasks and obtain the offset times for each word and phoneme uttered by the participant.

Due to synchronization problems between the audio recording and the recorded motion data, we first create a transcript of the entire recording by using WhisperX [37], an *Automatic Speech Recognition* (ASR) model, instead of aligning the recordings exclusively with the text of the verbal tasks. Another benefit of this approach is that we can also account for unforeseeable words that were possibly uttered at the beginning of the recording, and for which we did not have a transcript before. Then, we locate the verbal

Туре	Unified	Vive	Pico	Meta	Direction (*)
Facial	CheekPuff	Cheek_Puff_*	CheekPuff	CheekPuff*	left/right
	EyeClosed*	Eye_*_Blink	EyeBlink_*	EyesClosed*	left/right
	EyeLook*	Eye_*_*	EyeLook*	EyesLook*	left/right, down/up, in/out
	Jaw	Jaw_*	Jaw*	Jaw*	forward/thrust, left/right, open/drop
	LidTightener*	Eye_*_Squeeze	EyeSquint_*	LidTightener*	left/right
	UpperLidRaiser*	Eye_*_Wide	EyeWide_*	UpperLidRaiser*	left/right
	LipCornerDepressor*	Mouth_Sad_*	MouthFrown_*	LipCornerDepressor*	left/right
	LowerLipDepressor*	Mouth_Lower_Down*	MouthLowerDown_*	LowerLipDepressor*	left/right
	UpperLipRaiser*	Mouth_Upper_Up*	MouthUpperUp_*	UpperLipRaiser*	left/right
	LipCornerPuller*	Mouth_Smile_*	MouthSmile_*	LipCornerPuller*	left/right
	LipPucker*	Mouth_Pout	MouthPucker	LipPucker*	left/right
	LipSuckB	Mouth_Lower_Inside	MouthRollLower	LipSuck*B	left/right
	LipSuckT	Mouth_Upper_Inside	MouthRollUpper	LipSuck*T	left/right
	Mouth*	Mouth_*_*	Mouth*	Mouth*	lower/upper
	TongueOut	Tongue_LongStep*	Mouth*	Mouth*	lower/upper
Eye	LookDirection*	Gaze_Direction_*	LookDirection*	LookDirection*	X,Y,Z; left, right
	Position*	Gaze_Origin_MM_*	Position*	Position*	X,Y,Z; left, right
Head	DevicePosition*	DevicePosition*	DevicePosition*	DevicePosition*	X,Y,Z; left, right
	DeviceRotation*	DeviceRotation*	DeviceRotation*	DeviceRotation*	X,Y,Z,W; left, right

 TABLE 2. MAPPING FROM THE DEVICE-DEPENDENT MOTION DATA ATTRIBUTES TO THE UNIFIED DATA FORMAT. FOR N-TO-1 MAPPINGS FROM THE DEVICES TO THE UNIFIED FORMAT WE USE THE MEAN OF THE DIRECTIONS.

tasks in the transcript and correct any errors using the text of the specific task.

These transcriptions were then used as input for the *Montreal Forced Aligner* (MFA) [38], along with the full recordings. By being given the full transcriptions, the model accurately aligned them to the audio recordings and returned the offsets of when each word and phoneme was uttered.

As a last step, we had to convert the alignment offsets in the audio recordings to the actual timestamp ranges in the motion data files. To do this effectively, we interpolated the start and stop timestamps of the text tasks in the data with the start and stop alignment offsets of the same text tasks obtained from MFA. As a result, we could segment the text task data into word and phoneme segments.

#### 5.7. Data Availability

In total, we recorded 259 sessions. 19 of these sessions were missing one headset recording, resulting in a total of 499 individual headset recordings. Table 3 provides an overview of the number of samples segmented as described above. The dataset will be published alongside this paper.

# 6. Evaluation

Here, we present the evaluation that we performed on the dataset. Our main goal is to investigate the types of privacy inferences that can be made from facial motion data. However, we also perform the same experiment on eye gaze and head motion data to allow for comparison. First, we present the experiments we performed. Next, we detail the methodology for the biometric recognition system. Lastly, we present the results of the experiments.

#### 6.1. Experiments

We first want to learn if identification from facial motion data collected with MR headsets is possible. The prior work on facial motion videos (see Section 2.1) and on eye gaze identification (see Section 2.3) suggests that this should be possible.

In our first Experiment **E1**, we wish to learn if individuals can be identified from their facial motion data. For this experiment, we will investigate the identification for each headset separately, as well as all headsets together. We then investigate the influence of the sessions for Experiment **E2**. Hence, we use the first two sessions for training the biometric recognition system and then only test on the third session. This way we can see if the identification is stable over time. Next, we examine in Experiment **E3** if we can reidentify individuals when they start using different headsets, giving us insights how dependent the identification is on the headset type and if it can be generalized across MR headset types.

Besides identification, the related work (see Section 2.2) suggests that it should be possible to infer the facial expression and therefore we expect that it is possible to infer the emotion displayed in the non-verbal tasks.

TABLE 3. OVERVIEW OF THE DATASET REGARDING THE AMOUNT OF SAMPLES IT COMPRISES.

Segmentation	Total	Per Group		Per Device			Per Session		
Segmentation	Total	Α	В	Vive	Pico	Meta	0	1	2
Recordings	499	232	267	132	127	240	229	150	120
Tasks	19296	8883	10413	5175	4905	9216	8136	6750	4410
Words	197255	88477	108778	45087	51631	100537	82814	67362	47079

In Experiment E4, we test how good we can recognize the emotions displayed in our non-verbal tasks. Further, we also test if we can correctly classify which verbal task was performed. In Experiment E5, we investigate if the MR headset type can be inferred from the data collected. All headset data has the same format due to the unified data format, however, we expect that it is easy to infer which headset is being used due to device specific quirks. Then, we look at the inference of sensitive attributes about the user of the MR headset in Experiment E6. Here, we seek to infer the sex, English level, and personality trait of the user.

Lastly, we perform two experiments to better understand the identification from facial motion data. In Experiment E7, we perform the identification only on the verbal tasks or only on the non-verbal tasks to see which task type works better for identification. And in Experiment E8, we test how good we can identify individuals when we combine the facial motion data with the eye gaze and head motion data.

#### 6.2. Data Preparation & Splitting

For our evaluation, we use task-level segmentation of our dataset in the unified data format. We filter out the recordings performed on January 22, 2025, as some of the Vive's facial motion data values are missing. We then remove the timestamp column from the remaining samples and resample each one to 100 frames, normalizing the size of all samples.

Next, we split the data into training and testing datasets for the biometric recognition model. The testing dataset is used exclusively to calculate the model's final performance. Since different experiments require different data splits, we use multiple splits:

**Random:** For the random split type, we randomly split all samples, allocating 80% to the training dataset and 20% to the testing dataset.

**Session:** For the sessions split-type, we use the recordings from the first two sessions from each participant as the training dataset and the last session as the testing dataset.

Leave-one-headset-out-per-participant (LHPP): The LHPP split type uses two MR headsets per participant for the training dataset and one MR headset for the testing dataset. This allows the biometric recognition model to learn to recognize specific participants and to use data from each MR headset type.

**Participant:** The participant split type allocates 80% of participants to the training dataset and 20% to the test dataset. This type of split is used for attribute inference

experiments to prevent cross-contamination of the results, e.g. the model learning to recognize attributes by identifying the specific participant.

#### 6.3. Biometric Recognition Models

For our experiments, we use three different machine learning models as a biometric recognition system. The first is a simple fully connected neural network that receives each sample as a single vector. This neural network consists of at least two fully connected linear layers and a variable number of hidden layers, which are determined via hyper parameter optimization. After each linear layer, we use a Rectified Linear Unit (ReLu) activation function, as well as a dropout layer, to prevent overfitting. The second model is a Long Short-Term Memory (LSTM) that processes each sample frame-by-frame. To determine the most likely class, we first use a linear layer to reduce the size of the output vector to the number of classes. The third model is EYKT [17], a DenseNet-based architecture. Between each convolution block, the network uses batch normalization and the ReLu activation function. All networks use log softmax to perform the final classification step.

The training dataset is randomly split into a main training dataset and a validation dataset for model training. The main training dataset contains 90% of the data, and the validation dataset contains 10%. Each model is trained for a maximum of 100 epochs with early stopping if the validation accuracy does not increase for 10 epochs. We use negative log likelihood loss as the loss function and 1280 samples as the batch size.

We determine the best model parameters for each experiment by performing parameter optimization for 100 steps. See Table 4 for the optimized parameters. After optimization, we use the model with the best performance on the validation dataset and run it on the testing dataset to determine the final accuracy for each experiment.

TABLE 4. OVERVIEW OF THE OPTIMIZED PARAMETERS

Parameter	Range	Note
Layer Size	10-256	Only Simple & LSTM
Hidden Layers	0-2	Only Simple & LSTM
Learning Rate Step Size	10-100	All
Learning Rate Alpha	0.01-1	All
Optimizer Learning Rate	0.0001-0.1	All
Weight Decay	0.00001-0.01	All

Data Type	Model	Vive	Pico	Meta	All	Chance
Facial	Simple	0.78	0.83	0.63	0.68	0.02
	LSTM	0.69	0.77	0.59	0.58	0.02
	EYKT	0.94	0.98	0.88	0.9	0.02
	Simple	0.87	0.7	0.45	0.54	0.02
Eye	LSTM	0.86	0.59	0.47	0.49	0.02
	EYKT	1.0	0.87	0.92	0.78	0.02
	Simple	0.99	0.88	0.76	0.7	0.02
Head	LSTM	0.94	0.78	0.76	0.8	0.02
	EYKT	1.0	0.95	0.98	0.95	0.02

TABLE 5. IDENTIFICATION ACCURACY USING A RANDOM SPLIT

#### 6.4. Implementation

We implemented the biometric recognition models using Python (3.12) and PyTorch (2.6.0). As learning optimizer, we used Adam, and for the parameter optimization we used Optuna (4.3). The code used for our evaluation will be published alongside this paper.

#### 6.5. Results

Here, we present our evaluation results. As a metric, we always use the accuracy, which is defined as the correct classifications divided by all classifications. Further, we also always give the percentage of the largest class in the experiment-specific data split as the chance level.

For our Experiment **E1** (see Table 5), we used the random split to gain general understanding of how well the identification works. For the facial data, we find that the Pico achieves the highest identification of 98%, the Vive achieves 94%, the Meta achieves 88%, and using all headsets together we achieve 90%. This fulfills our expectation that identification on facial motion data is possible. Comparing to the eye and head data types, we find that for both we achieve 100% identification for the Vive and the EYKT model. In general, we can observe that the identification works for all data types, and all headsets, with the head motion data performing the best in general, though the facial and eye motions are not far behind.

Moving on to Experiment E2 (see Table 6), we now split the data according to their sessions into training and testing dataset. In general, we can see for the face data that all headsets and model combinations exceed the chance level for identification. The best result is 43% balanced accuracy for the face data of the Meta when using the EYKT model. We conclude that the identification across sessions is possible, but most of the learned features from E1 identify the specific session and are not general for the individual. Comparing E1 and E2 results, it is also interesting to see that in E2 the Meta performs far better for facial motions, while in E1 it has the worst performance of all three headset types.

Next, we test if we can recognize participants across different MR headsets in Experiment E3 (see Table 7). The

TABLE 6. IDENTIFICATION ACCURACY USING A SESSION SPLIT

Data Type	Model	Vive	Pico	Meta	All	Chance
Facial	Simple	0.11	0.26	0.29	0.2	0.04
	LSTM	0.11	0.11	0.24	0.17	0.04
	EYKT	0.14	0.23	0.43	0.27	0.04
Еуе	Simple	0.14	0.03	0.03	0.06	0.04
	LSTM	0.12	0.02	0.07	0.05	0.04
	EYKT	0.1	0.06	0.1	0.09	0.04
	Simple	0.0	0.02	0.16	0.08	0.04
Head	LSTM	0.0	0.01	0.15	0.06	0.04
	EYKT	0.0	0.01	0.16	0.07	0.04

TABLE 7. IDENTIFICATION ACCURACY USING THE LHPP SPLIT FOR ALL HEADSETS

Data Type Model	Simple	LSTM	EYKT	Chance
Facial	0.61	0.52	0.63	0.02
Eye	0.45	0.47	0.65	0.02
Head	0.49	0.46	0.63	0.02

LHPP split leaves for every participant one headset type for which the model has not seen any data, hence, we simulate that the user switches to a new type of MR headset. The best accuracy for facial motion data is achieved by EYKT with 63%, showing that identifying individuals across headsets is possible, however, at a lower rate than in our baseline E1.

In our Experiment **E4**, we tested emotion recognition using only emotion tasks (see Table 8). As expected, facial motion data was the most effective for emotion recognition, with 86% accuracy. However, eye and head motions also enabled some emotion recognition, with accuracy rates of 59% and 58%, respectively.

Since we have multiple devices, we tested whether we could identify which headset was used to record the data for Experiment E5 (see Table 9). Unsurprisingly, we can

 TABLE 8. EMOTION RECOGNITION ACCURACY USING A

 PARTICIPANT-WISE SPLIT FOR ALL HEADSETS

Data Type Model	Simple	LSTM	EYKT	Chance
Facial	0.86	0.86	0.86	0.33
Eye	0.45	0.33	0.59	0.33
Head	0.49	0.32	0.58	0.33

TABLE 9. DEVICE TYPE RECOGNITION ACCURACY USING A PARTICIPANT-WISE SPLIT FOR ALL HEADSETS

Data Type Model	Simple	LSTM	EYKT	Chance
Facial	1.0	1.0	1.0	0.48
Eye	1.0	1.0	1.0	0.48
Head	1.0	1.0	1.0	0.48

TABLE 10. VERBAL TASK RECOGNITION ACCURACY USING A
PARTICIPANT-WISE SPLIT FOR ALL HEADSETS

Data Type	Simple	LSTM	EYKT	Chance
Facial	0.78	0.89	0.96	0.17
Eye	0.56	0.6	0.68	0.17
Head	0.17	0.17	0.47	0.17

TABLE 11. ENGLISH LEVEL RECOGNITION ACCURACY USING A PARTICIPANT-WISE SPLIT FOR ALL HEADSETS

Data Type Model	Simple	LSTM	EYKT	Chance
Facial	0.72	0.73	0.69	0.73
Eye	0.73	0.73	0.71	0.73
Head	0.73	0.73	0.68	0.73

achieve 100% recognition accuracy for all data types.

In addition to recognizing emotions, we test whether the text task can be identified from the recorded data (see Table 10). The best recognition accuracy of 96% is again achieved using facial motion data.

We examine the results of the attribute inferences tested in Experiment **E6**. Table 11 shows the results for English level recognition, and Table 12 shows the results for classifying whether someone is an ambivert, extrovert, or introvert. For both attributes, the results are close to the level of chance, so we do not believe they can be inferred from the data. For sex recognition, shown in Table 13, there appears to be some information which can be extracted. Since the EYKT model achieved significantly less than chance level, and with only two classes (everyone identified as either male or female) in the dataset, we can simply invert the labeling.

To better understand which task type is better for identifying individuals, we ran identification Experiment **E7** on only the verbal and non-verbal tasks. See Tables 14 and 15 for a comparison. Our results show that verbal tasks perform better than non-verbal tasks. However, it also does not

TABLE 12. PERSONALITY RECOGNITION ACCURACY USING A PARTICIPANT-WISE SPLIT FOR ALL HEADSETS

Data Type Model	Simple	LSTM	EYKT	Chance
Facial	0.43	0.49	0.41	0.49
Eye	0.44	0.47	0.41	0.49
Head	0.49	0.49	0.38	0.49

 TABLE 13. Sex recognition accuracy using a participant-wise

 Split for all headsets

Model Data Type	Simple	LSTM	EYKT	Chance
Facial	0.65	0.81	0.67	0.81
Eye	0.72	0.81	0.58	0.81
Head	0.73	0.81	0.56	0.81

#### TABLE 14. IDENTIFICATION ACCURACY USING A RANDOM SPLIT FOR ONLY VERBAL TASKS FOR ALL HEADSETS

Data Type Model	Simple	LSTM	EYKT	Chance
Face	0.72	0.65	0.93	0.01
Eye	0.52	0.56	0.83	0.01
Head	0.79	0.77	0.95	0.01

TABLE 15. IDENTIFICATION ACCURACY USING A RANDOM SPLIT FOR ONLY VERBAL TASKS FOR ALL HEADSETS

Data Type Model	Simple	LSTM	EYKT	Chance
Facial	0.63	0.47	0.8	0.01
Eye	0.38	0.33	0.75	0.01
Head	0.64	0.41	0.89	0.01

appear that a reliable sex recognition can be implemented with facial motion data for now.

Lastly, we further investigated the identification potential of the data we collected. In Experiment **E8** (see Table 16), we tested the identification accuracy using all data types simultaneously. We found that combining the three data types increased identification accuracy to 99%, thereby outperforming the best all-headset result from Experiment E1 (see Table 5).

# 6.6. Summary of Results

- We are able to show that persons can be identified from their facial motions.
- The identification across different sessions is possible, however, the achieved accuracy is not on a level that is usable for any real-world system at the moment.
- We are able to show identification across different MR headset types.
- We are able to infer the displayed emotion, spoken text, and used MR headset.
- We are not able to infer the personality and spoken language of a person.
- For sex recognition, we see some indication that some sex-related information is contained in facial motion data.

# 7. Discussion

Facial motion data is a behavioral biometric factor that can be used for identification, so it should be treated as

 
 TABLE 16. Identification accuracy using a random split for all headsets

Model Data Type	Simple	LSTM	EYKT	Chance
Facial + Eye + Head	0.89	0.83	0.99	0.01

such when sharing it online. However, our results indicate that facial motion might not be stable enough to reliable identify individuals over long periods of time. Only larger studies with longer intervals between sessions can determine whether facial motion data poses a long-term privacy threat to individuals. We expect MR headsets to improve their ability to record facial motion data in the future, so we also expect privacy problems with facial motion data to increase.

Our text recognition results show that we can infer which text was spoken. This suggests that lip reading may be possible using facial motion data. Uncareful sharing of facial motion data might lead to the revelation of the content of private conversations, for example, when a person has muted themselves but is still sharing their facial motion data.

When we compare our eye gaze and head motion results to those of previous studies, such as GazebaseVR [19] for eye gaze data and Nair et al. [31] for VR data, we find that our identification results are are not as good, especially when considering multiple sessions. We believe this is because the tasks in our dataset are designed primarily to capture facial motion data. For example, GazebaseVR uses specific eye-tracking tasks, such as following a dot with one's eyes or reading tasks. In contrast, we only record data after participants read the tasks and push the button to start recording; therefore, we do not expect much eye motion during recording. Additionally, none of our tasks require head motion, so little variance is expected.

#### 8. Conclusion

In this paper, we present the first large-scale dataset of abstract face motions captured using MR headsets. The dataset contains multiple sessions, and each participant is recorded using multiple headset types. Using this dataset, we demonstrate that facial motion data is a privacy-sensitive behavioral biometric factor that can be used to identify individuals with up to 98% not considering sessions and 43% when considering sessions. Furthermore, we demonstrate that individuals can be identified even when using a new type of MR headset that the attacker has not seen before.

#### Acknowledgments

Funded by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy – EXC 2050/1 – Project ID 390696704 – Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of Technische Universität Dresden; Funded by Helmholtz Association, topic "46.23 Engineering Secure Systems". The studies were conducted in the Karlsruhe Decision&Design Lab (KD<sup>2</sup>Lab), an experimental laboratory funded by the DFG and the Karlsruhe Institute of Technology (INST\_12138411-1\_FUGG).

# References

 S. Hanisch, E. Muschter, A. Hatzipanayioti, S.-C. Li, and T. Strufe, "Understanding person identification through gait," *Proceedings on Privacy Enhancing Technologies*, 2023.

- [2] P. Cheng and U. Roedig, "Personal voice assistant security and privacy—a survey," *Proceedings of the IEEE*, vol. 110, pp. 476–507, Apr. 2022.
- [3] J. L. Kröger, O. H.-M. Lutz, and F. Müller, What Does Your Gaze Reveal About You? On the Privacy Implications of Eye Tracking, pp. 226–241. Springer International Publishing, 2020.
- [4] L. Benedikt, D. Cosker, P. L. Rosin, and D. Marshall, "Assessing the uniqueness and permanence of facial actions for use in biometric applications," *IEEE Transactions on Systems, Man, and Cybernetics* - Part A: Systems and Humans, vol. 40, pp. 449–460, May 2010.
- [5] J. Zhang and R. B. Fisher, "3d visual passcode: Speech-driven 3d facial dynamics for behaviometrics," *Signal Process.*, vol. 160, p. 164–177, July 2019.
- [6] R. E. Haamer, K. Kulkarni, N. Imanpour, M. A. Haque, E. Avots, M. Breisch, K. Nasrollahi, S. Escalera, C. Ozcinar, X. Baro, A. R. Naghsh-Nilchi, T. B. Moeslund, and G. Anbarjafari, "Changes in facial expression as biometric: A database and benchmarks of identification," in 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 621–628, IEEE, May 2018.
- [7] G. Moreira, A. Graca, B. Silva, P. Martins, and J. Batista, "Neuromorphic event-based face identity recognition," in 2022 26th International Conference on Pattern Recognition (ICPR), (Montreal, QC, Canada), pp. 922–929, IEEE, Aug. 2022.
- [8] T. Kopalidis, V. Solachidis, N. Vretos, and P. Daras, "Advances in facial expression recognition: A survey of methods, benchmarks, models, and datasets," *Information*, vol. 15, p. 135, Feb. 2024.
- [9] Z. Zhao, Q. Liu, and F. Zhou, "Robust lightweight facial expression recognition network with label distribution training," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 3510– 3519, May 2021.
- [10] Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: Multi-head cross attention network for facial expression recognition," *Biomimetics*, vol. 8, p. 199, May 2023.
- [11] Y. Chen, J. Wang, S. Chen, Z. Shi, and J. Cai, "Facial motion prior networks for facial expression recognition," in 2019 IEEE Visual Communications and Image Processing (VCIP), IEEE, Dec. 2019.
- [12] W. Zhang, X. Ji, K. Chen, Y. Ding, and C. Fan, "Learning a facial expression embedding disentangled from identity," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6755–6764, IEEE, June 2021.
- [13] J. P. Lee, H. Jang, Y. Jang, H. Song, S. Lee, P. S. Lee, and J. Kim, "Encoding of multi-modal emotional information via personalized skin-integrated wireless facial interface," *Nature Communications*, vol. 15, Jan. 2024.
- [14] X. Chen and H. Chen, "Emotion recognition using facial expressions in an immersive virtual reality application," *Virtual Reality*, vol. 27, pp. 1717–1732, Nov. 2022.
- [15] D. Lohr, H. Griffith, S. Aziz, and O. Komogortsev, "A metric learning approach to eye movement biometrics," in 2020 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–7, IEEE, Sept. 2020.
- [16] L. Friedman, M. S. Nixon, and O. V. Komogortsev, "Method to assess the temporal persistence of potential biometric features: Application to oculomotor, gait, face and brain structure databases," *PLOS ONE*, vol. 12, p. e0178501, June 2017.
- [17] D. Lohr and O. V. Komogortsev, "Eye know you too: Toward viable end-to-end eye movement biometrics for user authentication," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 3151–3164, 2022.
- [18] M. H. Raju, D. J. Lohr, and O. V. Komogortsev, "Evaluating eye movement biometrics in virtual reality: A comparative analysis of vr headset and high-end eye-tracker collected dataset," 2024.
- [19] D. Lohr, S. Aziz, L. Friedman, and O. V. Komogortsev, "Gazebasevr, a large-scale, longitudinal, binocular eye-tracking dataset collected in virtual reality," *Scientific Data*, vol. 10, Mar. 2023.

- [20] W. Shao, S. Luo, and Z. Yan, "Cross-content user authentication in virtual reality," in *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, ACM MobiCom '24, (New York, NY, USA), pp. 2098–2105, ACM, Dec. 2024.
- [21] S. M. Asish, A. K. Kulshreshth, and C. W. Borst, "User identification utilizing minimal eye-gaze features in virtual reality applications," *Virtual Worlds*, vol. 1, pp. 42–61, Sept. 2022.
- [22] A. Liu, L. Xia, A. Duchowski, R. Bailey, K. Holmqvist, and E. Jain, "Differential privacy for eye-tracking data," in *Proceedings of the 11th* ACM Symposium on Eye Tracking Research & Applications, ETRA '19, (New York, NY, USA), pp. 1–10, ACM, June 2019.
- [23] X. Ren, J. Fan, N. Xu, S. Wang, C. Dong, and Z. Wen, "Dpgazesynth: Enhancing eye-tracking virtual reality privacy with differentially private data synthesis," *Information Sciences*, vol. 675, p. 120720, July 2024.
- [24] E. Wilson, A. Ibragimov, M. J. Proulx, S. D. Tetali, K. Butler, and E. Jain, "Privacy-preserving gaze data streaming in immersive interactive virtual reality: Robustness and user experience," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, pp. 2257– 2268, May 2024.
- [25] E. Arabadzhiyska, O. T. Tursun, K. Myszkowski, H.-P. Seidel, and P. Didyk, "Saccade landing position prediction for gaze-contingent rendering," ACM Transactions on Graphics, vol. 36, pp. 1–12, July 2017.
- [26] Z. Hu, S. Li, C. Zhang, K. Yi, G. Wang, and D. Manocha, "Dgaze: Cnn-based gaze prediction in dynamic scenes," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, pp. 1902–1911, May 2020.
- [27] D. Ding, Z. Cao, Z. Gu, H. Chen, C. Qi, and F. Dong, "Foanet: Focus of attention prediction for foveated pre-rendering to enable high-quality edge vr," ACM Transactions on Sensor Networks, Mar. 2025.
- [28] M. R. Miller, F. Herrera, H. Jun, J. A. Landay, and J. N. Bailenson, "Personal identifiability of user tracking data during observation of 360-degree vr video," *Scientific Reports*, vol. 10, Oct. 2020.
- [29] J. Liebers, P. Horn, C. Burschik, U. Gruenefeld, and S. Schneegass, "Using gaze behavior and head orientation for implicit identification in virtual reality," in *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*, VRST '21, pp. 1–9, ACM, Dec. 2021.
- [30] A. G. Moore, R. P. McMahan, H. Dong, and N. Ruozzi, "Personal identifiability and obfuscation of user tracking data from vr training sessions," in 2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), pp. 221–228, IEEE, Oct. 2021.
- [31] V. Nair, W. Guo, J. Mattern, R. Wang, J. F. O'Brien, L. Rosenberg, and D. Song, "Unique identification of 50,000+ virtual reality users from head & hand motion data," in *32nd USENIX Security Symposium* (USENIX Security 23), SEC '23, (USA), pp. 895–910, USENIX Association, 2023.
- [32] P. Ekman and F. W.V., "Facial action coding system," *Environmental Psychology & Nonverbal Behavior*, 1978.
- [33] M. Horizon, "Face tracking for movement sdk for unity," 2025.
- [34] C. M. University, "Facs facial action coding system," 2001.
- [35] I. J. S. 37, "Information technology Vocabulary Part 37: Biometrics," standard, International Organization for Standardization, Geneva, CH, Feb. 2017.
- [36] L. Lu, J. Yu, Y. Chen, H. Liu, Y. Zhu, Y. Liu, and M. Li, "LipPass: Lip Reading-based User Authentication on Smartphones Leveraging Acoustic Signals," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, (Honolulu, HI), pp. 1466–1474, IEEE, Apr. 2018.
- [37] M. Bain, J. Huh, T. Han, and A. Zisserman, "Whisperx: Time-accurate speech transcription of long-form audio," in *INTERSPEECH 2023*, ISCA, aug 2023.
- [38] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *Proc. Interspeech 2017*, pp. 498–502, 2017.