Deep Motion Masking for Secure, Usable, and Scalable Real-Time Anonymization of Ecological Virtual Reality Motion Data

Vivek Nair* UC Berkeley
Wenbo Guo† Purdue University
James F. O'Brien‡ UC Berkeley
Louis Rosenberg§ Unanimous AI
Dawn Song¶ UC Berkeley

ABSTRACT

Virtual reality (VR) and "metaverse" systems have recently seen a resurgence in interest and investment as major technology companies continue to enter the space. However, recent studies have demonstrated that the motion tracking "telemetry" data used by nearly all VR applications is as uniquely identifiable as a fingerprint scan, raising significant privacy concerns surrounding metaverse technologies. In this paper, we propose a new "deep motion masking" approach that scalably facilitates the real-time anonymization of VR telemetry data. Through a large-scale user study ($N = 182$), we demonstrate that our method is significantly more usable and private than existing VR anonymity systems.

1 INTRODUCTION

The recent resurgence of research and development investiture into virtual reality (VR) and "metaverse" technologies has created an accelerated pace of technological improvements that are steadily making their way to consumer-facing VR devices. Newly announced products like the Apple Vision Pro [2] and Meta Quest 3 [19] blur the lines between virtual and augmented reality, resulting in extended reality (XR) systems that are expected to be more deeply and seamlessly integrated with our daily lives than ever before. Despite these changes, motion capture "telemetry" data remains fundamental to the operation of nearly all XR devices and applications.

While human motion patterns have been recognized as a uniquely identifiable and revealing biometric since at least the 1970s [4, 17], researchers are only beginning to understand the implications of this for motion data captured by XR devices. Recent studies have demonstrated that head and hand motion data captured by a VR device can be used to uniquely identify its user across a variety of applications [21, 37, 39], over long periods of time [20, 32], and at a rate of over 1 in 50,000 [28], comparable to that of a fingerprint scan [40]. Moreover, they show that a variety of potentially sensitive user data attributes can be inferred directly from VR telemetry streams [30]. Such results raise serious questions about whether XR devices can be used without involuntarily revealing a plethora of personal information to the device, application, and other XR users.

Researchers have proposed a number of methods for anonymizing VR motion data without unduly degrading the user experience [21, 25, 26]. However, current anonymization methods underestimate the identifiability of motion data when using sophisticated models trained on large datasets. In this paper, we propose "deep motion masking," a technique that uses deep learning to anonymize VR motion data more effectively than existing countermeasures.

Deep motion masking represents a multi-axis improvement over prior VR anonymization methods. Through a comprehensive evaluation, we demonstrate a $2.7\times$ improvement in the indistinguishability of anonymized motion data, and an over $20\times$ improvement in cross-session unlinkability. Our proposed system is capable of low-latency real-time anonymization of VR telemetry streams, making it concretely practical for deployment in new and existing VR systems.

Contributions:

- We propose a "deep motion masking" technique for scalable, real-time anonymization of VR telemetry data (§6).
- Using new and existing VR identification models, our evaluation ($N = 1,000$ users) shows at least a $20\times$ improvement in anonymity over prior VR privacy approaches (§7.1).
- Our large-scale usability study ($N = 182$ participants) demonstrates a nearly $3\times$ improvement in the indistinguishability of resulting anonymized motion data (§7.2).

2 BACKGROUND

VR systems use a variety of input and output devices to create an immersive visual, auditory, and haptic experience for users. In addition to being used for its intended purposes, the data generated by VR device sensors can be used adversarially to infer private user information. The SoK of Garrido et al. [8] provides a standard information flow and threat model for VR privacy research. We briefly recount the threat model and information flow of Garrido et al. to position our work within the landscape of VR privacy research.

2.1 Information Flow

A typical VR system sold today includes one head-mounted display (HMD) and two hand-held controllers. At a rate of between 60 and 144 times per second, the VR device measures the position and orientation of each of these three devices in 3D space (with six degrees of freedom), creating a "telemetry stream." These measurements are typically generated using a combination of inertial measurement units (IMUs) and either onboard cameras (also known as "inside-out" tracking) or external tracking stations (known as "outside-in" tracking). Many modern VR devices contain a number of additional sensors, such as LIDAR arrays, microphones, cameras, eye tracking, and body tracking systems. However, the focus of this paper is on the basic head and hand motion telemetry data that remains universal and fundamental to nearly all VR devices and applications.

In a typical VR system, telemetry data is generated by the VR device hardware and is then consumed by a VR application via an API provided by the device's firmware. The VR application uses this data to render frames to be displayed on the VR device, as well as to generate auditory and haptic stimuli for the user. In the case of a multi-user or "metaverse" application, the telemetry data is also forwarded to a server, which in turn forwards the data to other users in order to render a virtual representation (or "avatar") of the user on the devices of other users in the same virtual environment.

2.2 Threat Model

Because each entity in the above information flow (namely, the VR hardware, the application, the server, and another user) has access to the motion data stream of a target user, they could all potentially misuse such data in order to infer private user information. As such,

*e-mail: vcn@berkeley.edu
†e-mail: henrygwb@purdue.edu
‡e-mail: job@berkeley.edu
§e-mail: louis@unanimous.ai
¶e-mail: dawnsong@berkeley.edu

DOI 10.1109/VRW62533.2024.00096

they are all considered potential adversaries in the Garrido et al. threat model. Our emphasis in this paper is on protecting the motion data visible to external adversaries, namely VR game servers and other VR users. These adversaries are considered "weaker" in the Garrido et al. threat model, meaning that attacks available to them are typically available to all other adversaries. Moreover, attacks performed by these adversaries are generally the hardest to detect due to their remote and decentralized nature.

3 Related Work

Security and privacy in XR is a rapidly growing area of research that is summarized well by a number of existing survey and position papers [8, 31]. In this section, we summarize the body of work most directly relevant to this paper. We begin by detailing the history of motion-based biometrics and describe a number of studies illustrating relevant attacks on VR motion data. We then outline the relatively small number of proposed countermeasures to said attacks in comparison with the defensive system proposed herein.

3.1 VR Attacks

Prior work researching the privacy consequences of VR motion data specifically may be broadly categorized into identification studies, which use VR motion data to uniquely identify VR users, and profiling studies, which instead attempt to infer specific attributes such as age and gender. Many papers have analyzed the possibility of motion-based identification in VR [21, 24, 28], which are summarized well by the SoK papers of Stephenson et al. [38] and Garrido et al. [8]. A large number of studies have independently concluded that the head and hand motion data captured by VR devices is capable of accurately identifying users in a variety of applications.

A second major class of VR motion privacy research investigates profiling specific user attributes from head and hand movement patterns. For example, Tricomi et al. [39] use eye tracking data in addition to head and hand motion to accurately infer the gender and age of about 35 VR users. More recently, in a study of 1,006 VR users, Nair et al. [30] demonstrated that over 40 personal attributes, ranging from background and demographics to behavioral patterns and health information, can be accurately and consistently inferred from VR motion data alone. Additionally, multiple studies have demonstrated that adversarially designed VR applications can harvest further user data than passive observation alone [1, 27].

In summary, while motion data is an essential part of most VR experiences, prevailing research indicates that sharing this data with third parties carries significant security and privacy consequences. Thus, it is imperative to develop systems that enhance the privacy of VR motion data without impeding essential application functionality.

3.2 VR Defenses

There are a number of fundamental challenges that complicate the development of privacy-preserving mechanisms for VR motion data. First, there is a lack of fine-grained access control, as the exact same telemetry data that is necessary to provide legitimate multi-user functionality can also be used for adversarial purposes. While related work proposes access control for environmental data in XR systems [13, 15], the equivalent does not yet exist for motion data. Thus, instead of eliminating access to the VR motion stream, the data must somehow be transformed such that potential adversarial uses are thwarted while legitimate functionality remains intact. We compare this objective to a real-time voice changer that makes a speaker's voice unrecognizable while preserving spoken content and producing natural-sounding speech [6].

Further complicating attempts to protect the privacy of VR motion data is the need for any resulting defensive system to be real-time and low-latency. In many cases, even slight delays in a VR rendering pipeline can result in a phenomenon known as "VR sickness" [18]. This means that any realistic countermeasure must be fast and respect causality (i.e., cannot use future data to process past data). By contrast, VR attackers can be slow and non-causal, using an entire session of motion data at once to conduct their attack, creating a fundamental imbalance between attacker and defender capabilities in VR. Many adjacent research areas, such as gait recognition, lack these constraints. As such, proposed defenses in related domains are not necessarily directly applicable to VR.

With respect to VR motion data specifically, Miller et al. [21] have suggested a motion transmission method that only communicates joint rotation data, resulting in a 75% reduction in identifiability. Similarly, Moore et al. [25] suggest a method that transmits velocity data rather than positions, observing a 57% reduction in identification accuracy. On the contrary, Rack et al. [32] actually recommend the use of body-relative velocity and acceleration for identification purposes, citing an increase in identification accuracy rather than a reduction. MetaGuard [26], the prior state-of-the-art system, anonymizes VR motion data by applying bounded Laplacian noise [12] to specific dimensions of the telemetry stream that correspond to identifiable anthropometrics like height and wingspan. As a result, the system satisfies ε-differential privacy [7] and theoretically achieves an optimal noise versus privacy trade-off.

We seek to improve upon the existing countermeasures for two major reasons. First, the ad-hoc nature of the dimensions selected for anonymization is unlikely to be scalable when additional tracked locations are introduced to the system. As full-body tracking systems are increasingly becoming the norm for new VR devices [2, 19], proposed defensive mechanisms should at least plausibly demonstrate the potential to scale to more than three tracked locations in the future. Furthermore, existing countermeasures did not anticipate the extent to which users may be identified from a reduced set of features given a sufficiently powerful model. In the following section, we describe the substantial dataset utilized in this study, which enabled significant improvements in defensive VR technologies.

4 Dataset

The primary source of motion capture data used in this paper is "Beat Saber," a popular VR rhythm game in which players use a pair of sabers held in each hand to slice flying blocks that represent musical beats. Beat Saber is split into a number of levels or "maps," which consist of an audio track (typically a song) and a series of in-game obstacles that players must accurately interact with to achieve a high score. A number of third-party leaderboard services, such as "BeatLeader" [33], have emerged to allow players to capture and share high-quality telemetry recordings from within Beat Saber. The recordings have been aggregated and anonymized in BOXRR-23 [29], a publicly available VR motion dataset that contains over 3.5 million VR motion capture recordings submitted by nearly 100,000 users between February 2022 and April 2023. We primarily use the public BOXRR-23 dataset in this paper to improve the transparency, reproducibility, and extensibility of this work.

5 Problem Statement

We present in this paper a new "deep motion masking" approach to motion anonymization, which we use to create an improved VR privacy system. The goals of our new system are as follows:

- **Anonymity**: The primary goal of the system is to prevent users from being identified based on their motion data. Specifically, we invoke the same notion of anonymity as used in MetaGuard [26], *cross-session unlinkability*; given motion data with known user identities in a first session, the adversaries relevant to this paper (see §2.2) should not be able to identify the same set of users using their anonymized motion data from a second session. As in MetaGuard, we assume that adversaries have no other means of linking participant identities across sessions, such as IP addresses.

- **Usability**: The system must not significantly degrade the user experience by anonymizing user motion data. Specifically, we target the strong notion of *indistinguishability* of anonymized motion data from unmodified VR motion data.
- **Scalability**: The anonymization system should comprehensively anonymize every axis of motion data without manually engineering a solution for each feature.
- **Interactivity**: The system should minimize the perceived impact of the anonymization process on the interaction of the user with objects in the virtual world.

With these properties in mind, we now describe our new proposal for a "deep motion masking" system.

6 METHOD

At a high level, our method involves decomposing the plausible variance of human motion sequences into action-related variance and user-related variance. For this purpose, we train an "action encoder" model, which learns an embedding for the action a user is taking while ignoring the user's identity, and a "user encoder" model, which learns an embedding for the user's identity while ignoring the action they are taking. We then train an "anonymizer" model that anonymizes motion sequences by changing their user embedding without changing their action embedding. Finally, we train a "normalizer" model to remove unwanted noise added by the anonymizer. Each of the models we describe was implemented in Keras [14] and trained using the Adam optimizer [16] with a diminishing learning rate scheduler and early stopping based on a validation set. For each training step, and throughout this paper, we provide benchmarking results in §A of the full paper.

6.1 LSTM Funnel Architecture

First, we begin by proposing a new deep learning architecture that aims to internally replicate the idea of summarizing one-second motion subsequences by using a combination of Long Short-Term Memory (LSTM) [11] and Multi-Layer Perceptron (MLP) [9] layers. Figure 1 illustrates how the proposed architecture may be used to identify VR motion sequences. The model receives as input a 30-second motion sequence normalized to 30 frames per second, thus containing 900 frames in total. Using an LSTM layer, each frame is converted into a 256-dimensional feature vector. Then, an average pooling layer combines each one-second (30-frame) subsequence into a 256-dimensional summary. Next, another LSTM layer combines the sequence of 30 256-dimensional summaries into a flat 256-dimensional embedding. Finally, a fully connected MLP layer with softmax activation produces a classification output, with optional additional dense layers in between.

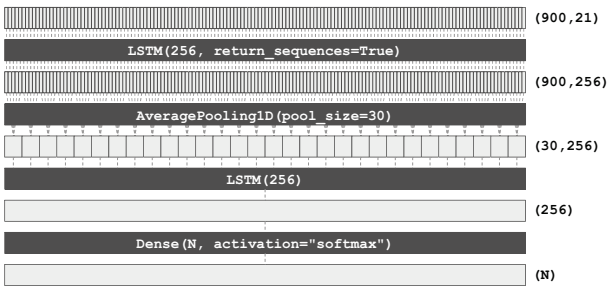

Figure 1: "LSTM funnel" identification architecture.

In essence, the architecture described above continues to represent VR motion sequences using summary statistics taken across one-second chunks, yet is able to outperform prior VR identification approaches for a few major reasons. First, instead of manually specifying summary statistics to be taken, such as mean, standard deviation, etc., the model is allowed to learn its own relevant statistics via the first LSTM layer. Second, instead of manually specifying how to summarize the classification of each subsequence, such as via a logarithmic sum of probabilities, the model is allowed to learn its own meta-classification method via the second LSTM and subsequent MLP layers. Moreover, the "featurization" and "classification" parts of the model are trained together in an end-to-end fashion, allowing the model to learn how to create complex statistics that result in optimal classification results. We term this approach the "LSTM funnel" architecture due to the dimensionality reduction performed by the average pooling layer. While fairly simple overall, to the best of our knowledge, this architecture has not yet been disclosed in general or has not been used for similar purposes.

6.2 Action Similarity

Next, we describe our method for measuring the similarity of the "action" performed in two separate VR motion sequences. To achieve this, we train an "action similarity" model using the architecture shown in Figure 2. The model is trained as a binary classifier that receives two 30-second telemetry sequences (900×21) as input. Each of the sequences is first passed through an identical encoder using the LSTM funnel architecture described in §6.1 to generate a 256-dimensional embedding. The Euclidean distance between these embeddings is then used to output a 1 if the two motion sequences correspond to the same action, and a 0 otherwise.

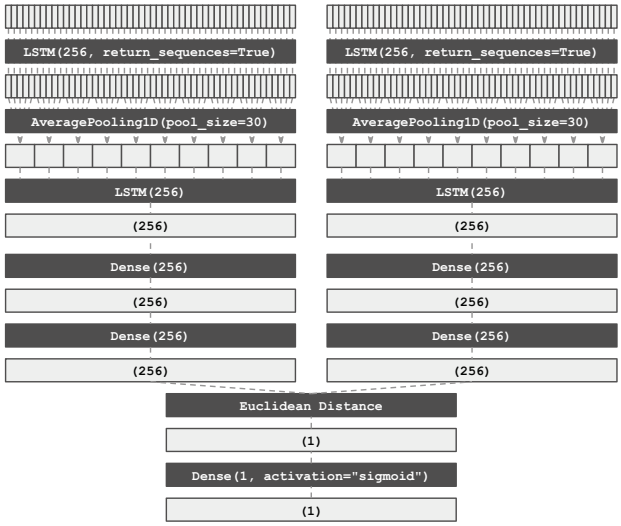

Figure 2: Siamese architecture for similarity models.

The approach illustrated in Figure 2 is sometimes known as a "Siamese neural network" [3]. Siamese architectures have previously been used in VR identification models [23], albeit with CNN layers rather than our LSTM funnel architecture. An advantage of this approach is that while it is trained as a binary classifier for "action similarity," a limb of the model can later be used on its own as an "action encoder," such that the Euclidean distance between two embeddings produced by the encoder reveals the similarity of actions in the inputs. To train the action similarity model, we randomly sampled 50,000 distinct pairs of "similar" motion sequences from the dataset of §4, and another 50,000 distinct pairs of "dissimilar" motion sequences. An additional 5,000 similar and 5,000 dissimilar pairs were sampled for validation, with a further 5,000 similar and

5,000 dissimilar pairs for testing. For the purpose of defining similarity, we use the "software.activity.id" attribute of the recordings provided in BOXRR-23 [29]. In this case, the attribute corresponds to the exact map the user is playing (see §4). In every instance, the two motion sequences constituting a pair of inputs originate from different users. The model is thus tasked to classify whether two different users are playing identical or different in-game levels.

When training the action similarity model on the 200,000 motion sequences (50,000 pairs $\times$ 2 classes) discussed above, early stopping occurred after the 156th epoch. The model achieved 100.00% training accuracy, 99.53% validation accuracy, and 99.40% testing accuracy. Therefore, we now have (1) a binary classifier that can determine with 99.4% accuracy whether two motion sequences correspond to the same map, and (2) an action encoder that has learned an approximate metric for measuring the similarity of two motions.

6.3 User Similarity

Next, we train a "user similarity" model, which is essentially the inverse of the action similarity model described above. Using the same architecture as before (Figure 2), we now randomly sampled 50,000 pairs of motion sequences from the same user, and another 50,000 distinct pairs of motion sequences from different users. Again, an additional 5,000 similar and 5,000 dissimilar pairs were sampled for validation, and 5,000 similar and 5,000 dissimilar pairs for testing. In every instance, the two motion sequences constituting a pair of inputs originate from different in-game maps. The model is thus now tasked to ignore the action and classify whether two motion samples originate from the same or different users.

When training the user similarity model on the 200,000 motion sequences (50,000 pairs $\times$ 2 classes) discussed above, early stopping occurred after the 27th epoch. The model achieved 97.94% training accuracy, 92.60% validation accuracy, and 92.81% testing accuracy. Therefore, in addition to the (1) action similarity and (2) action encoder models, we also have (3) a user similarity classifier that can determine with 92.8% accuracy whether two motion sequences correspond to the same user, and (4) a user encoder that has learned a metric for characterizing the user from a motion sequence.

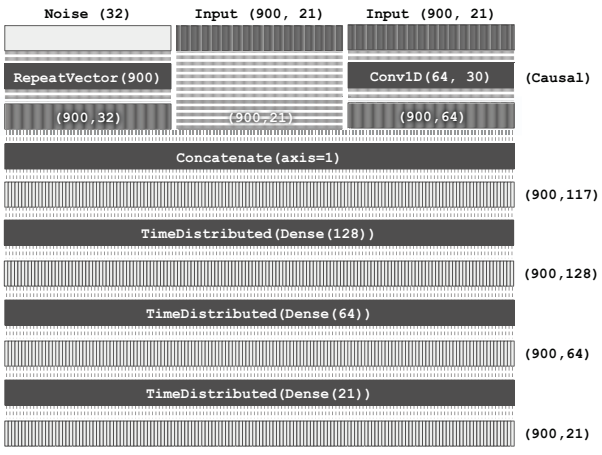

Figure 3: Architecture used for anonymizer model.

6.4 Anonymizer

Using the trained action similarity and user similarity models described above, we can now train the "anonymizer" model that performs the core deep motion masking functionality. The anonymizer model receives as input a 30-second motion telemetry sequence (900×21), and a 32-dimensional noise vector containing random Gaussian noise. It uses these values to output a corresponding 30-second motion sequence (900×21) that is an anonymized version of the input. Our anonymizer architecture is illustrated in Figure 3.

In addition to the motion input (900×21) and noise (32) (which is repeated to produce a (900×32) sequence), a learned 1D convolution (900×64) of the motion input is produced. These three sequences are then vertically concatenated to produce a (900×117) hybrid sequence. Multiple time-distributed dense layers are then used to reduce this sequence back to a (900×21) output sequence.

The intuition behind this architecture is that the dense layers effectively combine the noise and motion data to anonymize the motion data in a way that is consistent across each frame, creating a smooth and continuous motion output. This allows the motion to be anonymized in 3D space, but not across the time domain. Therefore, the 1D convolution is added to allow limited manipulation of time-series relationships in the data within a sliding one-second window.

Importantly, every component of this architecture respects causality; the model does not have the capability to "look into the future" when producing any output frame. For example, the 1D convolution uses causal padding such that only frames $N-30$ through N are used in the output of frame N. This allows the trained anonymizer model to be deployed in real-time on a frame-by-frame basis.

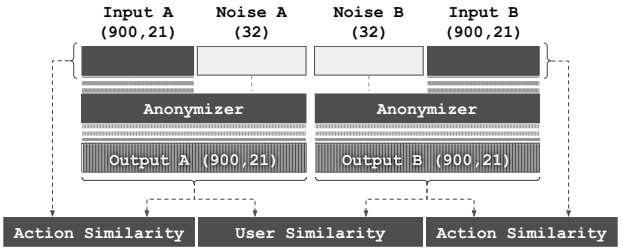

Figure 4: Siamese architecture for training anonymizer model.

Figure 4 shows how the action similarity and user similarity models are used to train the anonymizer model. First, the anonymizer is pre-trained for 20 epochs as an autoencoder with MSE loss, such that the output frames are initially nearly identical to the inputs, regardless of which noise values are provided. Then, a Siamese architecture is once again used. Leveraging the trained action and user similarity models (the weights of which are now frozen), the anonymizer is trained with the following loss function components:

1. The action embedding of input_A and output_A should always be as close as possible (irrespective of noise_A).
2. Similarly, input_B and output_B should always have as close of an action embedding as possible.
3. If $\mathsf{user}_A = \mathsf{user}_B$ and $\mathsf{noise}_A = \mathsf{noise}_B$, the user embedding for output_A and output_B should be as close as possible.
4. If $\mathsf{user}_A = \mathsf{user}_B$ and $\mathsf{noise}_A \neq \mathsf{noise}_B$, the user embedding for output_A and output_B should be far apart.

In other words, the action represented by an anonymized motion sequence should remain unchanged from the original motion sequence, helping to achieve the indistinguishability goal of our model. Furthermore, the intended use of the noise value is to be randomly sampled at the start of each new session, and then to remain consistent within that session. Thus, a user should assume a consistent faux identity within a session, but should assume distinct apparent identities across sessions, achieving cross-session unlinkability. Importantly, by using the adversarial training method in Figure 4, the anonymizer receives precise differentiable feedback from the action and user similarity models on how to achieve both of these goals.

An additional advantage of this training method is that it provides a tunable security parameter that can be used to adjust the balance of anonymity and usability while training the model. If additional usability is needed, more weight can be placed on loss components (1) and (2), causing the output motion to appear more similar to the input motion. On the other hand, if more anonymity is required, further weight can be put on loss components (3) and (4), emphasizing cross-session unlinkability of outputs. In our evaluation, we use equal weights for both components, meaning that indistinguishability and cross-session unlinkability are equally important goals.

To train the anonymizer, we randomly sampled 50,000 pairs of motion sequences, with both samples in any given pair coming from the same user. We then randomly sampled 50,000 pairs of random Gaussian noise vectors. For half of the pairs, the noise inputs are identical ($\mathsf{noise}_A = \mathsf{noise}_B$), while for the other half, they are different ($\mathsf{noise}_A \neq \mathsf{noise}_B$), per the loss function described above. An additional 5,000 pairs were sampled for testing. The model was trained for a full 500 epochs without early stopping.

The model achieved user similarity accuracy of 95.54% on the training data and 94.71% on the testing data. In other words, 94.71% of the time, the model correctly predicted that $\mathsf{user}_A = \mathsf{user}_B$ when $\mathsf{noise}_A = \mathsf{noise}_B$ and that $\mathsf{user}_A \neq \mathsf{user}_B$ when $\mathsf{noise}_A \neq \mathsf{noise}_B$. These numbers should be interpreted in light of the user similarity model's baseline accuracy of 92.81%. Importantly, on both datasets, the model achieved an action similarity accuracy of 100.00%; in every training and testing sample, the action similarity model correctly described the input and output motion as containing the same action.

6.5 Normalizer

While the anonymizer is effective at obscuring the identity of a VR user while keeping their big-picture actions looking the same, it introduces some undesirable noise to the telemetry signal (at the frame level) due to the lack of an incentive against doing so. One idea for combating this would be to use an adversarial architecture (e.g., GAN [10]) with a discriminator network that provides feedback to the anonymizer by attempting to distinguish anonymized motion from unmodified motion sequences. Unfortunately, we found this idea difficult to apply for our use case as discussed further in §8.2. Instead, we use a normalizer model that aims to reverse the effects of the anonymizer using the architecture in Figure 5.

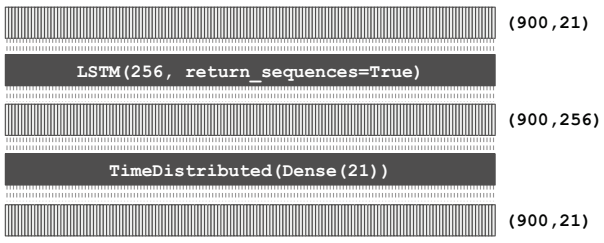

Figure 5: Normalizer model architecture.

The normalizer receives as input an anonymized motion sequence (900×21) and outputs a normalized motion sequence (900×21). The relatively simple architecture consists of an LSTM layer that returns a 256-dimensional state for each frame and a time-distributed dense layer that converts each state back to a 21-dimensional output. As with the anonymizer, the architecture obeys causality (e.g., no bidirectional layers) and can thus be deployed in a real-time setting.

To train the normalizer, we randomly sampled 50,000 motion sequences from random users and maps and anonymized each of them using random noise vectors. We then trained the normalizer using a subset of the anonymized motion sequences as inputs and the corresponding original motion sequences as the target outputs, with a mean squared error loss function. Using a portion of the sequences reserved for testing, we found that the mean squared error between input and output samples after z-score normalizing every dimension was reduced by about one order of magnitude.

Importantly, the normalizer model is not provided with the noise values used to anonymize the original motion sequences, and, during inference, does not have access to the original motion data. Therefore, it will never be able to fully recover the original motion sequences, and cannot reduce the anonymity of the motion sequences, as any deterministic algorithm that could undo the anonymization without access to the original motion or noise values could also be deployed by an adversary to defeat anonymized motion sequences. Instead, the normalizer network can only remove any component of the noise added by the anonymizer that is consistent or predictable across all anonymized motion sequences, which does not affect the actions or anonymity of any particular user.

The entire deep motion masking system architecture, with about 2.2 million parameters, is shown in §B of the full paper. Of these, 290k parameters are in the normalizer, with the action and user similarity models containing nearly one million parameters each. The anonymizer itself contains only about 65k trainable parameters, allowing it to run extremely quickly on its own.

6.6 Deployment

Deploying the trained models for post-hoc anonymization of motion recordings is now as simple as randomly sampling 32 Gaussian noise values, invoking the anonymizer model on the input motion sequence and noise values, and then running the normalizer model on the output of the anonymizer.

Based on our observations, we suggest a few simple optimizations to the above process. First, we observe that it is better for indistinguishability if the population mean and standard deviation of each motion dimension in anonymized recordings match the population mean and standard deviation of each motion dimension in unmodified motion. This population-level shift does not impact the anonymity of any individual user. Second, we recommend duplicating the first frame of motion 30 times before including the subsequent motion input. This ensures the 1D convolution buffer of the anonymizer model is always filled with real data, reducing apparent noise and instability in the first second of the anonymized output. Finally, the quaternions representing rotational dimensions of the output should be normalized to unit magnitude for validity.

The deep motion masking system can also be used in a real-time (streaming) setting. To do so, a buffer of the last 30 frames should be maintained and initially filled with 30 copies of the first frame. For each new frame, a corresponding anonymized frame can be produced by running the anonymizer's learned 1D convolution on the frame buffer, then concatenating its 64-dimensional output to the 21-dimensional input and 32-dimensional noise vector to produce a 117-dimensional hybrid vector. That hybrid vector can then be converted into a 21-dimensional anonymized output frame using the dense layers of the anonymizer. Next, the optional optimization of shifting the population mean and standard deviation of each motion dimension back to that of the general population can be applied. Finally, the resulting frame can be fed into the LSTM layer of the normalizer, and the 256-dimensional LSTM state can be used by the dense layer of the normalizer to recover a final 21-dimensional anonymized and normalized output frame. Again, the quaternions should be normalized to unit magnitude.

Overall, the real-time deployment of deep motion masking adds no delay other than the computational delay of invoking the anonymizer and normalizer models, which we found to be less than 1 ms. Due to the causal design of the architecture, the anonymized and normalized output in the streaming setting is identical to the result of the post-hoc anonymization process.

7 Evaluation

Having fully described our proposed deep motion masking approach, we now present a detailed evaluation of the privacy and usability of the resulting system. Our evaluation directly compares the cross-session unlinkability and indistinguishability of our system to that of MetaGuard [26], the prior state-of-the-art for VR motion privacy.

7.1 Anonymity

First, we analyze the impact of our deep motion masking system on cross-session linkability. If the system is effective at anonymizing VR motion data, it should be able to trick our LSTM funnel classification model (§6.1) into wrongly classifying anonymized users in most instances. However, to ensure that our anonymizer didn't overfit by only fooling our own classification model, we also include the Random Forest identification model of Miller et al. [21] and LightGBM-based identification model of Nair et al. [28].

Furthermore, we train each model both as an oblivious adversary, which is trained on unmodified motion sequences from each user and tested on anonymized motion sequences, and as an adaptive adversary, which is trained on anonymized motion sequences from within a session and tested from anonymized motion sequences in another session. Per our definition of cross-session unlinkability in §5, none of the models are trained on multiple independent sessions of anonymized motion, as we operate under the assumption that no external identifiers can be used to link sessions together.

To perform the evaluation, we randomly selected 1,000 users from the dataset of §4. In order to be representative of average VR users, we only include users for which between 30 and 100 recordings were present; about 20,000 such users exist in the dataset. For each user, we selected 10 recordings to constitute the first session (for training) and another 10 recordings to constitute the second session (for testing). We then anonymized either one or both sessions (depending on the type of adversary), using either MetaGuard or the full post-hoc anonymization pipeline detailed in §6.6. The results of training and testing each of the considered identification models on each set of data are summarized in Table 1 below.

	Miller et al. [22]		**Nair et al. [28]**		**LSTM Funnel (§6.1)**	
	Oblivious	*Adaptive*	*Oblivious*	*Adaptive*	*Oblivious*	*Adaptive*
Unmodified	90.3%	90.3%	91.0%	91.0%	96.5%	96.5%
MetaGuard [26]	57.4%	79.5%	67.0%	84.3%	81.3%	96.3%
DMM (§6)	1.5%	1.2%	3.1%	3.5%	3.7%	0.1%

Table 1: Identification accuracy for various adversaries and model architectures, with and without anonymization.

As demonstrated by the results of Table 1, deep motion masking is significantly better than MetaGuard at anonymizing users across sessions. While MetaGuard users remain up to 96% identifiable, deep motion masking reduces identification accuracy to less than 4%, representing a 20× to over 100× improvement in anonymity.

As expected, adaptive adversaries are usually better at identifying anonymized users across sessions, as information about what the user looks like when using the anonymity tool of choice (albeit with different noise values) can be incorporated into the identification model. In the case of MetaGuard, this allows the LSTM funnel architecture to perform at nearly full accuracy, as the model learns to ignore anonymized dimensions and identify users by the unmodified dimensions. Interestingly, however, the LSTM funnel model actually performs significantly worse with the deep motion masking samples when trained adaptively. This is likely because component (3) of the loss function used to train the anonymizer model (§6.4) is measured by a user encoder based on the LSTM funnel architecture. The anonymizer model therefore is particularly good at tricking the LSTM funnel architecture into learning fictitious user attributes and consequently becoming worse at identifying users.

7.2 Usability

To evaluate the indistinguishability of motion data anonymized with deep motion masking, we conducted a large-user study (N=182). The study consisted of an online survey in which users were asked to watch VR motion recordings from the game Beat Saber in the Beat Saber web replay viewer tool [35] after reading and agreeing to an informed consent document. Four treatments were tested:

1. As a negative control group, we included unmodified VR motion recordings from the dataset of §4 that will certainly be indistinguishable from natural human motion.
2. As a positive control group, we included completely AI-generated motion recordings created by CyberRamen [36], a machine learning model trained to play Beat Saber. As it stands, these recordings are easily distinguishable from natural motion, serving as a good test of response quality.
3. As a baseline treatment group, we included recordings anonymized using MetaGuard [26] with the "height," "wingspan," and "room size" defenses enabled at the "medium" privacy settings suggested by the authors.
4. As our new treatment group, we included recordings anonymized with deep motion masking using the same models and processes as the anonymity evaluation (§7.1).

Users were given one set of recordings at a time, consisting of four recordings of different users playing the same map in Beat Saber. To remove confounding variables, all recordings in all sets were first normalized to 30 FPS and trimmed to the first 30 seconds. One of the four recordings in each set was additionally treated (i.e., "anonymized") using one of the four treatments listed above. Each user was shown 12 such sets of recordings in a randomized order, corresponding to a slow, medium, and fast song for each of the four treatment groups described above. For each set, their task was to decide which (if any) of the four recordings was modified. To aid their decision, users could view each replay in slow motion, zoom in on particular areas, and view the motion from a variety of perspectives. When recruiting participants for our study, we focused primarily on finding VR users with significant Beat Saber experience, as such users are more likely to be familiar with what natural VR motion data should look like, and thus are likely to be more challenging and discerning critics of our system. With that in mind, we primarily recruited participants through social media pages related to VR, and through VR interest groups like CVRE [5]. However, we also wanted to ensure that some number of novice users participated in the study, and recruited a small number of participants from a broader general population for that purpose.

The study ran for two weeks, from September 20th, 2023 through October 3rd, 2023, and received 241 responses in that time. We removed the 59 responses that were either blank or answered all six of the control questions incorrectly, leaving 182 valid responses. Of those, 149 were from expert Beat Saber players (with 100 or more hours of in-game experience), and the remaining 33 participants were novices (with 0 to 100 hours of experience). Figure 6 shows the observed distinguishability for each of the evaluated treatments.

The negative control group has a surprisingly high rate of distinguishability in our results (18%). This indicates that when unsure about which replay was modified, users in our study were prone to randomly guessing one of the four replays rather than indicating that all four replays were unmodified. With that in mind, the replays anonymized with deep motion masking were only marginally more distinguishable than the negative control group. Moreover, deep motion masking represents a significant improvement over the MetaGuard [26] system, with nearly a 3× reduction in the rate of distinguishability, particularly for expert users. Using a standard χ^2 test, the difference between MetaGuard and deep motion masking is highly statistically significant with $p < 0.01$.

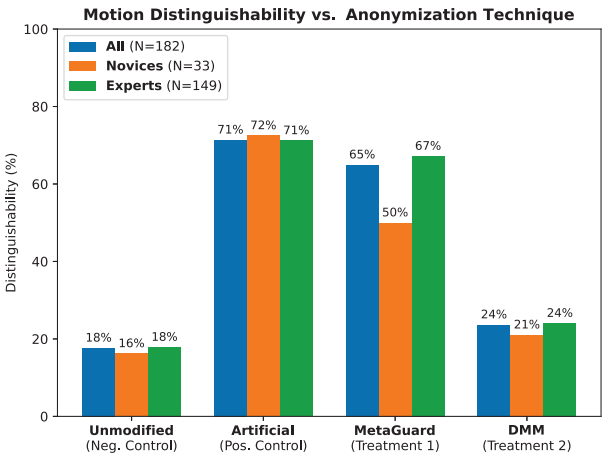

Figure 6: Results of user study on distinguishability (lower is better).

7.3 Ethics

The primary source of data for this study is the BOXRR-23 dataset [29], a publicly available dataset intended for use in VR research, including security and privacy research. This dataset has already been used in published research papers in the VR security and privacy domain [28]. It contains built-in privacy measures, such as pseudonymization of participants, and was reviewed by the legal and ethics boards of its authors prior to release. We specifically only use the BeatLeader part of the dataset in our research; these users agree to the use of their data for "research topics such as VR security, privacy, and usability" in the BeatLeader privacy policy [34].

Other than the BOXRR-23 dataset, the only additional data used in this paper is from our usability study in §7.2. All participants in the survey were adults over the age of 18, and no vulnerable populations were specifically targeted in this study. Participants consented to their inclusion in academic research by reading and agreeing to an informed consent document before proceeding in the survey. Users optionally provided their Beat Saber username, but no further identifiable information was collected. Information collected consisted exclusively of the users' selections of which recordings they believed were modified. Therefore, the likelihood of any harm to participants, either through participating or through a later breach of confidentiality, is exceedingly low.

All aspects of this study, including our use of the public BeatLeader data and our collection of survey responses in §7.2, were also independently reviewed and approved by an OHRP-certified IRB under protocol number 2023-06-16467.

8 Discussion

Anonymizing VR motion data inherently involves diverging from the original motion data to some extent. The approach detailed in §6 ensures that such deviations correspond mostly to apparent differences not in the actions being taken but rather in the user taking the actions. This results in the system being highly suitable for motion data intended for consumption by human observers, as demonstrated in §7.2.

On the other hand, there will always be VR applications in which very high precision is required, such as telemedicine, competitive e-sports, or remote operation of equipment. If anonymity is still desired in such an application, an alternative solution, such as secure multi-party computation or trusted execution environments, may be more suitable. Thus, we recommend a two-channel approach for VR motion data, with one system handling real-time anonymization of low-fidelity motion for human eyes, and another handling precise motion data for asynchronous computational use. Deep motion masking presents a secure, usable, and scalable solution for the former scenario, while the latter merits further investigation.

8.1 Limitations

One major limitation of our system is that it has only been trained on data from a single VR application, Beat Saber. This is because there are currently about four orders of magnitude more motion data available from Beat Saber than any other VR application, with deep learning models benefiting from large amounts of training data. Unlike prior work using this dataset, we don't allow our model to see anything specific to Beat Saber, such as block positions and timings. Therefore, it should be possible to train a deep motion masking model, using the present architecture, on motion data from any VR application, if enough motion data were available. However, without such data, we cannot confidently claim that the evaluation results will generalize to other applications.

Another major limitation of deep motion masking is that it loses the provable security properties of MetaGuard [26]. One of the most significant features of MetaGuard is that it obeys ε-differential privacy, and thus provides provable security and privacy properties. However, that provability only extends to the specific dimensions that the authors consider in the paper. This creates a weakness, as rotational dimensions are excluded entirely. Thus, while proving the security of our deep learning approach is significantly harder, the method empirically provides better cross-session unlinkability than MetaGuard as demonstrated in §7.1.

8.2 Future Work

One important area of future work in this field is extending motion anonymization systems support to full-body tracking data. Deep motion masking is particularly suitable for this purpose, as it doesn't involve manually engineering features between pairs of tracked objects, and may in fact be immediately applicable to full-body telemetry streams. At present, we lack a sufficiently large full-body motion capture dataset to use for training. However, as next-generation VR devices move towards full-body tracking, such data may become readily available, and the importance of full-body motion anonymization will simultaneously increase.

On the subject of data, future work may focus on procuring large-scale VR motion datasets from applications other than Beat Saber. Demonstrating the generalizability of deep motion masking to a wide variety of VR games and applications is an important step toward the potential adoption of such a system. Finally, we hope to see future work that explores various other architectures and techniques for masking VR motion data.

9 Conclusion

Deep learning is emerging as a powerful method for the usable real-time anonymization of sequential data (e.g., voice anonymization [6]). In this paper, we've shown that deep learning can also be an effective tool for anonymizing VR telemetry data by developing a technique we call deep motion masking, which is analogous to a real-time voice changer for movement patterns. By decomposing the space of motion variability into action-related variation and user-related variation, our model is effective at hiding user identity while maintaining action similarity, leading to better indistinguishability and cross-session unlinkability than prior methods.

Acknowledgments

We greatly appreciate the advice and support of Allen Yang, Ananya Kharche, Atticus Cull, Beni Issler, Bjoern Hartmann, Brandon Huang, Charles Dove, Chris He, Christian Rack, Dziugas Ramonas, Eric Paulos, James Smith, Mani Malek, Rui Wang, Shuixian Li, Viktor Radulov, Xiaoyuan Liu, and Zade Lobo. This work was supported in part by the Minderoo Foundation, Meta Reality Labs, the

National Science Foundation, the National Physical Science Consortium, the Fannie and John Hertz Foundation, and the Berkeley Center for Responsible, Decentralized Intelligence. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of their employers or the supporting entities. We sincerely thank all of the users who participated in our study or contributed to the BOXRR-23 dataset for making this work possible.

AVAILABILITY

The source code and documentation necessary to train and test all of the models and evaluations discussed in this paper are available on our GitHub repository under a BSD license:

`https://github.com/metaguard/metaguardplus`

Appendices

Additional sections and appendices are available in the preprint version of this paper: `https://arxiv.org/abs/2311.05090`.

REFERENCES

[1] N.-M. Aliman and L. Kester. Malicious design in aivr, falsehood and cybersecurity-oriented immersive defenses. In *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pp. 130–137, 2020. doi: 10.1109/AIVR50618.2020.00031

[2] Apple Vision Pro.

[3] D. Chicco. *Siamese Neural Networks: An Overview*, pp. 73–94. Springer US, New York, NY, 2021. doi: 10.1007/978-1-0716-0826-5_3

[4] J. E. Cutting and L. T. Kozlowski. Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the Psychonomic Society*, 9(5):353–356, May 1977. doi: 10.3758/BF03337021

[5] Collegiate VR Esports League (CVRE).

[6] J. Deng, F. Teng, Y. Chen, X. Chen, Z. Wang, and W. Xu. V-Cloak: Intelligibility-, naturalness- & Timbre-Preserving Real-Time voice anonymization. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 5181–5198. USENIX Association, Aug. 2023.

[7] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3):211–407, 2013. doi: 10.1561/0400000042

[8] G. M. Garrido, V. Nair, and D. Song. Sok: Data privacy in virtual reality. In *24th Privacy Enhancing Technologies Symposium (PETS 24)*, 2024.

[9] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[11] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[12] N. Holohan, S. Antonatos, S. Braghin, and P. Mac Aonghusa. The Bounded Laplace Mechanism in Differential Privacy. *Journal of Privacy and Confidentiality*, 10(1), Dec. 2019. doi: 10.29012/jpc.715

[13] S. Jana, D. Molnar, A. Moshchuk, A. Dunn, B. Livshits, H. J. Wang, and E. Ofek. Enabling Fine-Grained permissions for augmented reality applications with recognizers. In *22nd USENIX Security Symposium (USENIX Security 13)*, pp. 415–430. USENIX Association, Washington, D.C., Aug. 2013.

[14] Keras: Deep learning for humans.

[15] Y. Kim, S. Goutam, A. Rahmati, and A. Kaufman. Erebus: Access control for augmented reality systems. In *32nd USENIX Security Symposium*, pp. 929–946. USENIX Association, Aug. 2023.

[16] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] L. T. Kozlowski and J. E. Cutting. Recognizing the sex of a walker from a dynamic point-light display. *Perception & Psychophysics*, 21(6):575–580, Nov. 1977. doi: 10.3758/BF03198740

[18] J. J. LaViola. A discussion of cybersickness in virtual environments. *SIGCHI Bull.*, 32(1):47–56, jan 2000. doi: 10.1145/333329.333344

[19] Meta Quest 3: New Mixed Reality VR Headset.

[20] M. R. Miller, E. Han, C. DeVeaux, E. Jones, R. Chen, and J. N. Bailenson. A large-scale study of personal identifiability of virtual reality motion over time, 2023.

[21] 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*, 10(1):17404, Oct. 2020. Nature Publishing Group. doi: 10.1038/s41598-020-74486-y

[22] R. Miller, N. Banerjee, and S. Banerjee. Within-system and cross-system behavior-based biometric authentication in virtual reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 311–316, 03 2020. doi: 10.1109/VRW50115.2020.00070

[23] R. Miller, N. K. Banerjee, and S. Banerjee. Using siamese neural networks to perform cross-system behavioral authentication in virtual reality. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 140–149, 2021. doi: 10.1109/VR50410.2021.00035

[24] A. G. Moore, T. D. Do, N. Ruozzi, and R. P. McMahan. Identifying virtual reality users across domain-specific tasks: A systematic investigation of tracked features for assembly. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 396–404, 2023. doi: 10.1109/ISMAR59233.2023.00054

[25] 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, 2021. doi: 10.1109/ISMAR52148.2021.00037

[26] V. Nair, G. M. Garrido, and D. Song. Going incognito in the metaverse: Achieving theoretically optimal privacy-usability tradeoffs in VR. In *36th ACM Symposium on User Interface Software and Technology (UIST 23)*, 2023.

[27] V. Nair, G. M. Garrido, D. Song, and J. O'Brien. Exploring the Privacy Risks of Adversarial VR Game Design. In *23rd Privacy Enhancing Technologies Symposium*, 2023. doi: 10.56553/popets-2023-0108

[28] 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)*, pp. 895–910. USENIX Association, Aug. 2023.

[29] V. Nair, W. Guo, R. Wang, J. F. O'Brien, L. Rosenberg, and D. Song. Berkeley open extended reality recordings 2023 (BOXRR-23): 4.7 million motion capture recordings from 105,852 extended reality device users, 2023.

[30] V. Nair, C. Rack, W. Guo, R. Wang, S. Li, B. Huang, A. Cull, J. F. O'Brien, M. Latoschik, L. Rosenberg, and D. Song. Inferring private personal attributes of virtual reality users from head and hand motion data, 2023.

[31] V. Nair, L. Rosenberg, J. F. O'Brien, and D. Song. Truth in motion: The unprecedented risks and opportunities of extended reality motion data, 2023.

[32] C. Rack, K. Kobs, T. Fernando, A. Hotho, and M. E. Latoschik. Extensible motion-based identification of XR users using non-specific motion data, 2023.

[33] V. Radulov. BeatLeader.

[34] V. Radulov. BeatLeader Privacy Policy.

[35] V. Radulov, K. Ngo, D. F. Goberna, B. Bukaty, J. Kerrane, and J. Baron. Beat Saber web replays.

[36] D. Ramonas. CyberRamen.

[37] C. Schell, A. Hotho, and M. E. Latoschik. Comparison of Data Encodings and Machine Learning Architectures for User Identification on Arbitrary Motion Sequences. In *2022 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pp. 11–19, Dec. 2022. ISSN: 2771-7453. doi: 10.1109/AIVR56993.2022.00010

[38] S. Stephenson, B. Pal, S. Fan, E. Fernandes, Y. Zhao, and R. Chatterjee. Sok: Authentication in augmented and virtual reality. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 267–284, 2022. doi: 10.1109/SP46214.2022.9833742

[39] P. P. Tricomi, F. Nenna, L. Pajola, M. Conti, and L. Gamberini. You can't hide behind your headset: User profiling in augmented and virtual reality. *IEEE Access*, 11:9859–9875, 2023. doi: 10.1109/ACCESS.2023.3240071

[40] C. L. Wilson. Biometric accuracy standards, 2003.