MetaFL: Privacy-preserving User Authentication in Virtual Reality with Federated Learning

Ruizhi Cheng
George Mason University
Fairfax, VA, USA
rcheng4@gmu.edu

Yuetong Wu
Texas A&M University
College Station, TX, USA
yuetongw@tamu.edu

Ashish Kundu
Cisco Research
San Jose, CA, USA
ashkundu@cisco.com

Hugo Latapie
Cisco Research
San Jose, CA, USA
hlatapie@cisco.com

Myungjin Lee
Cisco Research
San Jose, CA, USA
myungjle@cisco.com

Songqing Chen
George Mason University
Fairfax, VA, USA
sqchen@gmu.edu

Bo Han
George Mason University
Fairfax, VA, USA
bohan@gmu.edu

Abstract

The increasing popularity of virtual reality (VR) has stressed the importance of authenticating VR users while preserving their privacy. Behavioral biometrics, owing to their robustness and ease of collection, compared to traditional modes such as passwords, have become a favored authentication choice. While current approaches that utilize behavioral biometrics to train classifiers for authentication yield promising accuracy, they cause privacy breaches by sharing sensitive data with a server to train a central model. In this paper, we present MetaFL, a first-of-its-kind privacy-preserving VR authentication framework that leverages federated learning (FL) on multi-modal motion data. The design of MetaFL is motivated by our key insight that *various modalities of motion data uniquely affect authentication performance for individual users and among different users*. It is attributed to the fundamental challenge of privacy-preserving user authentication: users can access only their own data with limited global knowledge. To tackle this issue, MetaFL judiciously selects the most suitable modalities for each user, which is decomposed into within-user ordering and between-user selection to eliminate the complex interplay between various conflicting factors. Moreover, we develop a personalized strategy to initialize FL models, further improving authentication accuracy. Our extensive performance evaluation on six public datasets shows that MetaFL outperforms state-of-the-art FL-based models (*e.g.,* 17–28% higher authentication accuracy), and its accuracy gap with the non-privacy-preserving central model is small (*i.e.,* only <2%).

CCS Concepts

• **Security and privacy** → **Authentication**; **Privacy protections**;
• **Human-centered computing** → **Virtual reality**.

Keywords

User Authentication, Biometrics, Privacy Preservation, Virtual Reality, Federated Learning

SENSYS '24, November 4–7, 2024, Hangzhou, China
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0697-4/24/11
https://doi.org/10.1145/3666025.3699322

ACM Reference Format:
Ruizhi Cheng, Yuetong Wu, Ashish Kundu, Hugo Latapie, Myungjin Lee, Songqing Chen, and Bo Han. 2024. MetaFL: Privacy-preserving User Authentication in Virtual Reality with Federated Learning. In *The 22nd ACM Conference on Embedded Networked Sensor Systems (SENSYS '24), November 4–7, 2024, Hangzhou, China.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3666025.3699322

1 Introduction

Extended reality (XR), which encompasses augmented, virtual, and mixed reality (AR/VR/MR), has gained increasing popularity, attributed to the immersive experiences it affords to users, especially in the burgeoning metaverse era [22]. As XR applications undergo rapid evolution, they pose significant challenges in safeguarding sensitive personal information, such as credit card details and medical records [107]. Thus, in this expanding landscape, it is imperative to design robust security measures, particularly in user authentication, to ensure strong protection against potential risks [87].

In the realm of XR, three types of authentication methods are applicable, each relying on distinct modes: knowledge (*e.g.,* passwords and PINs), physical biometrics, and behavioral biometrics. Traditional knowledge-based methods rely on *static* information, making them vulnerable to guesswork [11] and shoulder-surfing attacks [107]. Hence, the common practice includes frequent updates (*e.g.,* changing passwords every 90 days) [35] and two-factor authentication, which typically necessitates an additional device (*e.g.,* a smartphone) [30]. Additionally, these methods require cumbersome interactions for typing passwords or PINs on virtual keyboards in XR environments [87], leading to significant user inconvenience.

Authentication based on physical biometrics shares similar limitations as knowledge-based methods by relying on (quasi-)static information. Most physical biometrics are non-cancelable [79]. Once they are stolen or compromised, they cannot be reissued or revoked. Moreover, these methods may be intrusive for XR applications (*e.g.,* face recognition) and require additional expensive sensors [87], which are largely absent from commercial XR headsets [108]. In contrast, behavioral biometrics, such as head [51, 70], hand [56, 66], and gaze [57, 106] movements, are resilient against identity theft due to their dynamic nature [4, 6] and can be easily collected on existing off-the-shelf XR headsets [57], making them a favorable candidate for authenticating XR users [87]. We summarize the three authentication methods for XR users based upon existing literature reviews [69, 87, 107] in Table 1.

Methods	Robustness to Attacks	Convenience to Users	Additional Requirements
Knowledge	Low to Med.	Low	No
Physical	Low to Med.	Medium	Yes
Behavioral	High	High	No

Table 1: Comparison of three XR authentication methods.

Authentication based on behavioral biometrics typically trains classifiers for decision-making [91] by learning unique features from various *modalities*, which refer to the motion data of, for instance, the head or hand, captured from sensors on the headset [99]. However, prior studies utilize centralized training for authenticating XR users [62, 95, 108], which requires them to upload raw data to a central server, raising serious privacy concerns [69]. As a result, users may be unwilling to share their sensitive data due to fears of misuse or unauthorized surveillance [15]. On the other hand, federated learning (FL) [100] enables users to share only their model weights with the server and avoids exposing the raw data for training. Furthermore, during inference, users perform on-device authentication with trained models, eliminating the upload of sensitive data to the server and thus enhancing privacy safeguards.

Our key insights from a motivational study (§2.2) are that the significance of various modalities in authenticating VR users is not uniform, and the best modalities may vary for different users. However, state-of-the-art FL-based schemes, such as FedAwS [103] and FedUV [38], fail to account for the diversity of modalities. This oversight results in suboptimal performance for authenticating VR users, as the biometric data of each user is highly non-independent and identically distributed (non-IID) [109]. Our further investigations show that features of the best modality combination (*e.g.,* motion data of head and right hand) for a user should exhibit high density[1] with minimal noise and maintain a long distance from those of others in the feature space.

Inspired by these insights, we propose MetaFL, which is, to the best of our knowledge, the first privacy-preserving VR authentication framework that leverages FL on multi-modal motion data to boost authentication accuracy while minimizing information leakage. The novelty of MetaFL lies in its privacy-preserving selection of the optimal modality combination for each user *under extremely non-IID scenarios*, where the client has access to only a user's own data of a single class (*i.e.,* positive label). This challenge is unique to privacy-preserving user authentication tasks. Although modality selection has been explored by previous efforts in FL [97] and other privacy-preserving techniques such as multi-party computation (MPC) [54], they assume clients have both positive and negative labels to guide modality selection. Similarly, previous FL-based clustering approaches [21, 28, 29], designed to cluster or distinguish clients based on their data distributions, also rely on this assumption. As such, they are not applicable to MetaFL.

Our choice of leveraging FL to address users' privacy concerns in VR authentication is motivated by its practicality, as evidenced by the deployment in Google's speech recognition model [31] and the adoption for user authentication by Qualcomm [81] and Apple [34]. Although other orthogonal privacy-preserving techniques such as MPC [8] and fully homomorphic encryption (FHE) [9] can be potentially used for authentication, their practicality may still be poor. For instance, training an effective user authentication model often necessitates inter-party communication to exchange global information [91], rather than relying solely on local data, to reduce the false positive rate (FPR), one of the most important metrics in authentication systems [89]. However, FHE dramatically increases the data size after encryption (*e.g.,* over 3000× [98]), leading to substantial communication burdens [9], while MPC is inherently limited by its significant communication challenges [32]. In contrast, as we will show in §6.5, the communication overhead of MetaFL is manageable (*e.g.,* ~0.1 Mbps per user on average).

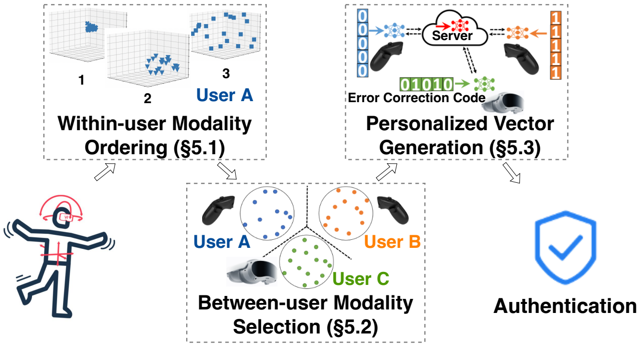

1
2
3
User A
Within-user Modality Ordering (§5.1)
00000
Server
11111
01010
Error Correction Code
Personalized Vector Generation (§5.3)
User A
User B
User C
Between-user Modality Selection (§5.2)
Authentication

Figure 1: Key system components of MetaFL.

The key challenge of designing MetaFL lies in how to efficiently select the optimal modality combination for each user without compromising privacy while making the extracted features from these modalities as distinguishable as possible. To address this challenge, MetaFL decomposes the problem by first intelligently ordering modality combinations for each user separately based on density. Then, it jointly selects the best modality combination for each user by considering both the density and the distance between them in the feature space. As shown in Figure 1, MetaFL consists of three major components with their respective challenges and solutions as follows.

Within-user Modality Ordering (§5.1). To select the best modality combination with high density, it is intuitive to filter out noise from the data before ordering. However, denoising high-dimensional motion data is computationally demanding on mobile VR devices. Moreover, the number of data samples may be reduced after denoising, which affects the density. Thus, the density before and after denoising should be meticulously balanced for effective ordering. In MetaFL, we first design a lightweight method to denoise data and then propose a novel density-calculation metric for users to order their modality combinations locally, which jointly considers the density before and after denoising.

Between-user Modality Selection (§5.2). To minimize misclassification probability, we should simultaneously optimize the density of chosen modality combinations for different users and the distance between them in the feature space. It presents a non-trivial endeavor since users are restricted from sharing sensitive information with each other and with the server, but modality selection introduces interdependence (*e.g.,* distance calculation) among different users' decisions. The brute-force method for this problem has an exponential time complexity. To address this challenge, MetaFL

[1] We will present a formal definition of density in §5.1.

takes an effective heuristic approach that first selects the modality combinations for a group of users based solely on density and then determines the most suitable modalities for others by considering both density and distance factors.

Personalized Vector Generation (§5.3). Capitalizing on our design of between-user modality selection, we propose a personalized vector generation method that creates a unique error correction code (ECC) for each user locally, aiming to increase the distance between users' ECCs. The generated ECCs, in turn, ensure a long distance among the features of selected modality combinations for different users without sharing sensitive information with third parties. Specifically, we introduce a tunable random-vector design to balance the tradeoff between privacy preservation, authentication performance, and inference latency.

Implementation of MetaFL and Performance Evaluation (§6). We implement a prototype of MetaFL and extensively evaluate its performance on six publicly available datasets. We summarize the experimental results as follows.

- MetaFL drastically improves authentication accuracy by 17–28%, compared to state-of-the-art FL-based models such as FedAwS [103] and FedUV [38], and has only a <2% gap with the non-privacy-preserving central model.
- MetaFL is capable of making an authentication decision in <250 ms, which is comparable to FedUV. However, it remarkably improves authentication accuracy by up to 25% when using the same setup of ECCs as FedUV.
- MetaFL achieves up to 76.2% bandwidth reduction compared to FedUV while maintaining higher authentication accuracy and reducing inference time by >300ms.
- When using data collected on different days for training and testing, the accuracy of MetaFL is <1% lower than that of five-fold cross-validation, demonstrating that the modality combinations selected by MetaFL for users are robust.

In addition to the above key results, our security and privacy analysis of MetaFL reveals its resilience to impersonation, mimicry, model inversion, and network-based attacks (§7). We also explore the potential of MetaFL in defending against other attacks and delve into various considerations regarding the real-world deployment of MetaFL (§8). Note that although we primarily focus on VR as a case study in this paper, our design is generic and can be applied to AR/MR as well, as AR/MR headsets can collect the same type of motion data for authentication (*e.g.,* head, hand, and gaze movements) [87]. This work does not raise any ethical issues.

2 Background and Motivation

2.1 Background

User Authentication in VR. Existing VR authentication systems are based on either knowledge or biometrics [87]. Knowledge-based methods such as passwords and PINs are inconvenient and vulnerable to security breaches in VR (*e.g.,* shoulder-surfing attacks) [108]. Thus, biometric-based schemes have gained prominence by offering better accessibility and robustness. Compared to physical-based ones, behavioral biometrics, such as body motion, have surged in popularity for user authentication in VR due to their resilience to identity theft [4, 6], ease of collection [57, 70, 80], and promising performance (*e.g.,* >95% accuracy) [65].

Biometric-based user authentication comprises two phases: enrollment and testing [41, 42]. During enrollment, user-specific biometric data is collected and used to train the authentication model, creating a unique user profile. In the testing phase, user input is compared with the stored profile to verify identity. Embedding-based classifiers [46] are commonly used in biometric-based user authentication [23, 103]. Let $g_\theta : x \to \mathbb{R}^D$ be a neural network that maps the input x (*e.g.,* raw motion data) to a D-dimensional input embedding vector $g_\theta(x)$, $w_y \in \mathbb{R}^D$ be the trained embedding vector of class y, and $d(\cdot)$ be a distance function. Suppose y is the true class of x, the loss function can be defined as [46]:

$$\mathcal{L} = d(g_\theta(x), w_y) - \min_{z \neq y} d(g_\theta(x), w_z) \quad (1)$$

It is designed to minimize the distance between the input embedding vectors and the true class embedding vector (positive loss, the first term to correctly authenticate users), while maximizing the minimum distance from the embedding vectors of others (negative loss, the second term to avoid misclassification). Thus, the training goal is to make the class embedding vectors of different users well separated in the feature space. A test user could pass authentication if the input embedding vector is sufficiently close to the class embedding vector of the claimed user.

To improve accuracy, VR authentication typically utilizes multiple sources of biometrics [65, 80, 87]. However, existing approaches require users to upload their biometrics to the central server for model training. We refer to these models as central models, which raise privacy concerns, for example, recognition of physical activities [75] and personal information leakage (*e.g.,* health data) [69]. Thus, there is a pressing need for designing privacy-preserving approaches when authenticating VR users with their biometrics.

FL for User Authentication. FL is a distributed learning framework using local data to train a global model by exchanging model updates instead of raw data [100]. Although leveraging FL for training authentication models may preserve user privacy [45, 53], it poses a unique challenge in that it can optimize only the positive loss in Eq. (1) since users can access only their own data of a single class. However, the negative loss is necessary to ensure effective training [10]. FedAwS [103] addresses the above issue by introducing a regularization to spread out class embedding vectors. However, it requires users to reveal their embedding vectors to the server, which may still lead to privacy leakage [45].

To address the privacy concern, FedUV [38] leverages ECCs to represent class embedding vectors, with the following loss function:

$$max\left(0, 1 - \frac{1}{c} v_y^T W g_\theta(x)\right) \quad (2)$$

where c, v_y, and W denote the length of the ECC, the ECC, and the matrix that projects v_y to the embedding vector of class y, respectively. The server and clients jointly train and exchange $W \in \mathbb{R}^{c \times D}$ using FedAvg [64]. The loss function of FedUV involves only the positive loss because it proves that minimizing the positive loss (*i.e.,* $Wg_\theta(x) = v_y$) also minimizes the negative loss. Thus, users can use their ECCs to train the model solely with the positive loss, avoiding revealing their embedding vectors to the server.

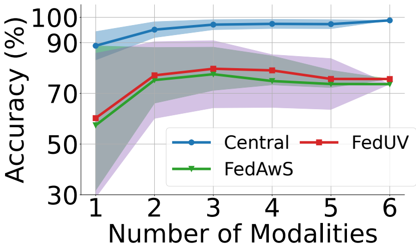


Figure 2: The accuracy of central, FedUV, and FedAwS models using different numbers of modalities. The bands represent 95% confidence intervals (CIs).

FedAwS and FedUV achieve high accuracy (*i.e.,* close to that of central models) for traditional scenarios (*e.g.,* authentication with face or handwriting), which are not commonly used in VR [87].

2.2 Motivation

To explore the efficacy of FedAwS and FedUV for authenticating VR users, we evaluate their performance with the dataset released by Miller *et al.* [65], which has been widely used in the research of VR authentication [66, 67]. It contains the motion data of 41 users' heads and hands when they throw a virtual ball 20 times in VR. We refer to this dataset as **`Throw`**. As the baseline, we train a central model with a loss function based on the softmax cross-entropy [60]. We train all models with five-fold cross-validation and calculate the authentication accuracy on the validation set for each fold, defined as the proportion of true positive and true negative predictions out of all predictions [89]. The structure and hyperparameter setting of these models are the same as that in §6.1. For FL-based models, we follow the original setup of FedAwS and FedUV, utilizing all available data from each client for training. The accuracy is 73.6% (SD = 3.5%), 75.7% (SD = 2.7%), and 98.2% (SD = 1.1%) for FedAwS, FedUV, and the central model, respectively.

To understand the gap between FL-based models and the central one, we adopt a fine-grained approach to define modalities, by considering the position and orientation of each body part as distinct modalities. The same is applied to data originating from different body parts (*e.g.,* left hand, right hand, and head). This is based on the consideration that some modalities may exhibit random patterns, which can negatively affect authentication accuracy. For example, a user's non-dominant hand may move randomly during tasks performed by the dominant hand. Thus, we investigate the impact of various modalities on the accuracy of FL-based authentication, which is underexplored, by training the models with different modality combinations. The **`Throw`** dataset includes six modalities: the position and orientation of the headset and both controllers, leading to a total of $\binom{6}{1} + \cdots + \binom{6}{6} = 63$ combinations.

Figure 2 shows the averaged accuracy with different numbers of modalities. The accuracy of FedAwS and FedUV varies drastically with different numbers of modalities. In contrast, the accuracy of the central model increases with the number of modalities. We further find that the best modality combination that leads to the highest accuracy may vary for different users. Thus, we train FedUV and FedAwS with the best modality combination(s) for each user. The average accuracy increases to 90.7% and 88.3% for FedUV and FedAwS, respectively. These findings call attention to revisiting FL-based models for VR user authentication.

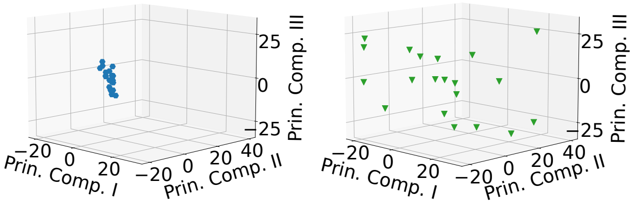


Figure 3: Visualization of motion data of the right controller position (left) and data of all six modalities (right) for user #19 after applying principal component analysis (PCA). Using only the right controller position achieves 100% accuracy, while combining all six modalities obtains 0% accuracy.

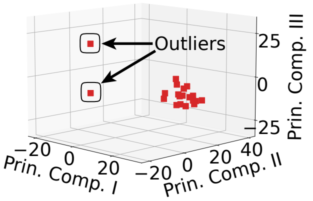


Figure 4: Visualization of data of the best modality combination for user #39 after applying PCA. Due to the noise, the models achieve 100% accuracy only three times (out of five).

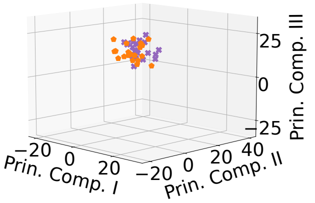


Figure 5: Visualization of data of the highest-density modality combinations for users #11 and #34 after applying PCA. They are misclassified due to the overlap of their data points.

Key Insights. To investigate the differences between various modalities and visualize the high-dimensional motion data [75], we apply principal component analysis (PCA) [44] on the 20 traces of each modality combination for all users. We use PCA for visualization, instead of other dimension-reduction techniques such as t-SNE [94], to ensure consistency with our proposed methods (§5.1). Specifically, we reduce the data dimension to 3D. We verify that it can still explain >70% variance, preserving enough information [36]. We have the following key insights.

#**1**. *The features of the best modality combinations tend to exhibit high density.* For example, Figure 3 shows the results of the best modality combination for user #19 that contains only the position of the right hand (left) and using all six modalities (right). The data points in the left sub-figure are better concentrated (*i.e.,* they have a high density) than those in the right one, which are scattered. As a result, training FedAwS and FedUV with data of this best modality combination achieves much higher accuracy (100%) for this user than training them with all six modalities (0%).

#**2**. *The noise in motion data of a modality combination leads to unstable authentication accuracy for FL-based models.* For instance, Figure 4 shows the results of the best modality combination for user #39. Most of the data points aggregate well, with the exception of two outliers. As a result, training FedAwS and FedUV with this modality combination achieves 100% accuracy only three out of five times for this user. The accuracy of others is under 85%. The reason is that low-quality data, such as noise, can significantly degrade the performance of FL models [33, 61].

#**3**. *If each user selects the modality combination with the highest density but the features of modality combinations selected by different users are close to each other in the space, the trained FL model may misclassify them.* For example, Figure 5 shows the results of the modality combinations with the highest density for users #11 and #34. The two groups of features overlap with each other. As a result, the learned class embedding vectors for them are close to each other in the feature space, leading to misclassification when training FedUV and FedAwS with these modalities.

Summary. Our analysis shows that blindly utilizing all modalities in FL-based authentication models leads to suboptimal performance. It highlights the common issue of applying existing FL-based authentication models for VR users: they do not consider the diversity of various modalities. This issue is particularly important in FL-based models where users have access to only their own data, hindering them from learning unique features. Nevertheless, the goal of authentication is to differentiate different class embedding vectors, which requires joint optimization among users.

3 Threat Model

In devising our threat model, we consider three potential adversaries: clients, the server, and network-based attackers.

Client-based Attacks. We assume the existence of a formidable adversary who possesses ample resources and time to execute comprehensive attacks. Given the portability of VR headsets, they may be stolen or lost, leading to unauthorized usage. Moreover, unauthorized access could be made by individuals acquainted with the device owner [62]. Considering these factors, we focus on the following two most commonly encountered threat models in behavioral biometric-based authentication [87]. Additionally, we discuss how MetaFL can potentially defend against other attacks in §8.

#**1**. *Impersonation Attacks:* Attackers use their own data to gain system access, possibly after a device is lost.

#**2**. *Mimicry Attacks*: Attackers observe a legitimate user's movements and attempt to replicate them to gain unauthorized access.

Server-based Attacks. During FL model training, the server may try to infer clients' private information from their updates. We assume the server could execute *gradient-based inversion attacks* [40], where it uses uploaded gradient updates to infer client data. Meanwhile, the server could conduct *model inversion attacks* [24], attempting to infer users' raw data from embedding vectors.

Network-based Attacks. These attackers could attempt to intercept communication between clients and the server. Even when data is encrypted, adversaries can still exploit side-channel information, such as WiFi signals [3] or packet patterns [88], to infer sensitive behavioral biometrics.

4 MetaFL Overview

MetaFL is an innovative privacy-preserving VR authentication framework with FL. Although Figure 2 shows that leveraging multiple modalities can potentially improve authentication accuracy, the key challenge of MetaFL lies in how to effectively select optimal modality combinations for all users without privacy leakage.

To address it, MetaFL first judiciously orders modality combinations based on density for each user (as indicated by Insights #1 and #2 in §2.2). Then, it selects the most suitable combinations by considering both density and distance in the feature space (Insight #3 in §2.2), as a locally optimal selection may not be globally optimal. The rationale behind such a design is that determining the optimal modality combinations for all users requires global information (*e.g.,* calculating feature distances across users). By decomposing the problem into the above two parts, we can minimize the exchanged information between users and the server during modality selection. Furthermore, benefiting from our modality-selection approach, MetaFL generates personalized initial class embedding vectors, enhancing accuracy by increasing their distances in the feature space. In the inference stage, clients record the movements of test users and then employ locally stored models to authenticate them.

Besides our proposed modality selection in MetaFL, we could also choose to optimize existing FL-based authentication models such as FedAwS and FedUV by directly handling low-quality data from some modalities. However, the improvements may be closely tied to a specific FL model, and different models may require different modifications. Thus, MetaFL adopts a generic approach, treating the FL-based authentication model as a black box, making it applicable to any existing model.

5 System Design of MetaFL

5.1 Within-user Modality Ordering

Problem. To decompose the selection of optimal modalities for all users, which requires global information, we first address the problem of ordering different modality combinations for each user based on their density and whether they include noisy data. Such ordering is determined by each user independently and will be used for selecting the best modality combinations for all users in §5.2.

Challenges. Although similar problems have been studied in anomaly detection, they are computationally intensive by focusing on high-dimensional data [14] and thus are not suitable for VR headsets. On-device ordering of modality combinations for each user requires MetaFL to devise a lightweight scheme. Moreover, the amount of training data is a critical factor affecting classification models' performance [46]. As some input data may be removed during denoising, naively ordering modality combinations with their density after denoising may result in suboptimal performance.

Our Approach. To design a lightweight scheme for denoising modality combination data and calculating density, we first apply PCA [44] to reduce the data dimension to 3D, which still preserves enough information (§2.2). As PCA is a linear approach, it is more efficient [76] than other methods such as t-SNE [94]. After that, we employ the mean shift clustering algorithm [19] in the resulting 3D space to denoise the data. Mean shift is a lightweight and practical approach for clustering low-dimensional data [27]. By grouping data points based on their proximity to one another, mean shift helps us identify isolated points, which can be considered noise.

Simply relying on the data samples after denoising to calculate the density for ordering modality combinations may not be optimal. The reason is that denoising may decrease the number of data samples. A small number of training data samples can degrade the performance of classification models [46, 96]. More importantly, it may affect the density of the modality combination and lead to inappropriate ordering that impacts authentication accuracy, as we

	FedUV			FedAwS		
Setup	D+O	O	-	D+O	O	-
Acc.	84.9%	80.2%	75.7%	82.2%	77.3%	73.6%

Table 2: The accuracy of FedUV and FedAwS for the Throw dataset, when each user selects the modality combination with the highest density with and without denoising and uses all modalities (D: denoising and O: ordering).

will demonstrate in §6.7. To address this issue, we jointly consider the number of data samples of modality combinations before and after denoising when calculating their density.

We now introduce a formal definition of the density for a modality combination. Let $\mathcal{S}^i = \left(x_i^1, \cdots, x_i^n\right)$ be the n input motion data samples of this modality combination for user i. The logical center of their embedding vectors V_i can be calculated as [82]:

$$V_i = \frac{1}{n} \sum_{j=1}^{n} g_\theta(x_i^j) \tag{3}$$

Let the maximum distance between V_i and the n input embedding vectors be $R_i = max\ (d(V_i, g_\theta(x_i))$. In embedding-based classification, a hypersphere with its center at V_i and its radius being R_i could represent the feature space of $\mathcal{S}^i$ [59, 77]. We denote the curvature of $\mathcal{S}^i$ as $\mathbb{C}$ (*i.e.,* $\mathbb{C} = 1/R_i$). Thus, a smaller radius R_i (*i.e.,* larger $\mathbb{C}$) implies a potentially higher density, as data samples will be more closely aggregated.

For a modality combination M, let $\mathbb{C}_b$ ($\mathbb{C}_a$) and S_b (S_a) be the curvature and the number of data samples of M before (after) denoising, respectively. Its density is defined as the weighted average of the curvature before and after denoising:

$$\mathbb{D}_M = \mathbb{C}_b \times \frac{S_b}{S_b + S_a} + \mathbb{C}_a \times \frac{S_a}{S_b + S_a} \tag{4}$$

After ordering modality combinations based on their density, we can simply select the highest-density one for each user. Table 2 demonstrates the effectiveness of within-user modality ordering (with and without denoising) for improving the accuracy of FedUV and FedAwS on the Throw dataset. Applying within-user modality ordering with denoising can improve the accuracy of FedUV and FedAwS by ~9% compared to training them with data from all six modalities (75.7% for FedUV and 73.6% for FedAwS, as shown in Figure 2). Moreover, denoising can improve the accuracy by ~5% for both FedUV and FedAwS compared to without denoising.

5.2 Between-user Modality Selection

Problem. Solely relying on density for modality selection is suboptimal, as suggested by Insight #3 in §2.2. The reason is that if we select the modality combination for each user independently by considering only density, the features extracted from its input embedding vectors may be too close to those of others in the feature space (Figure 5), leading to misclassification. Therefore, we should jointly examine the density and distance factors to select the optimal modality combinations for users.

Challenges. Optimal modality selection requires maximizing the distance between input embedding vectors of the modality combinations chosen by different users. Those vectors will be eventually represented by the class embedding vectors learned by the FL model, which leads to a long distance between the input embedding vector of a given user and others' class embedding vectors and improves authentication accuracy. One way to achieve this goal is to share the class embedding vectors of all users with the server, which can separate them during training (as proposed by FedAwS). However, doing this will lead to privacy leakage [38, 45]. Thus, it is non-trivial to simultaneously optimize the density and distance factors for selecting optimal modality combinations for users without sharing sensitive data such as class embedding vectors with the server.

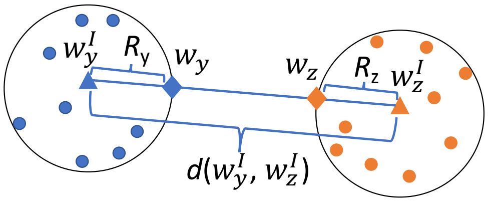


Figure 6: Illustration of the minimum distance between different class embedding vectors. Suppose w_y^I and w_z^I are two initial embedding vectors for classes y and z. The minimum distance between the final vectors after training is $d(w_y^I, w_z^I) - R_y - R_z$, which is achieved when they are at w_y and w_z, respectively (*i.e.,* on the boundary of the feature space).

Our Approach. To maximize the minimum distance between class embedding vectors, we propose to initialize each user's class embedding vector as the mean of all input embedding vectors for the selected modality combination before training. Typically, class embedding vectors are randomly initialized[2] from a Gaussian distribution [55]. However, this may lead to the initialized vector staying outside the feature space of input embedding vectors [2], resulting in the gradient vanishing problem [72].

Conversely, our initialized vector is a linear combination of all input embedding vectors, making it a proper starting point. Although the class embedding vector may change during the training process, it is guaranteed to lie in the same feature space finally [2]. Thus, as illustrated in Figure 6, we can calculate the minimum distance between the final embedding vector for class y and that of other classes as $\rho_y = \min_{y \neq z} \left(d(w_y^I, w_z^I) - R_y - R_z\right)$, where w^I is the initialized class embedding vector. $\rho_y \leq 0$ implies that the feature spaces of different classes overlap, indicating a high misclassification probability. In other cases, after knowing ρ_y, we can calculate the upper bound of the misclassification probability for class y.

Proposition 1. Let the expected distance between an input embedding vector $g_\theta(x)$ to its true class embedding vector w_y be $\tau = \mathbb{E}(d(g_\theta(x), w_y))$. Then, $\tau \leq 2R_y$ (*i.e.,* the diameter of the hypersphere after denoising). Given that the curvature of the hypersphere after denoising $\mathbb{C}_a$ should not be smaller than that before denoising $\mathbb{C}_b$ (*i.e.,* $\mathbb{C}_a \geq \mathbb{C}_b$), Eq. (4) indicates that $\mathbb{C}_a \geq \mathbb{D}_M$. The misclassification probability of class y satisfies

$$\begin{aligned} &P\left(\exists z \neq y\ s.t.\ d(g_\theta(x), w_y) \geq d(g_\theta(x), w_z)\right) \\ &\overset{i}{\leq} \frac{2\tau}{\rho_y} \leq \frac{4R_y}{\rho_y} \overset{ii}{=} \frac{4}{\mathbb{C}_a \rho_y} \leq \frac{4}{\mathbb{D}_M \rho_y} \end{aligned} \tag{5}$$

where $\overset{i}{\leq}$ is based on Markov's inequality and $\overset{ii}{=}$ is based on the definition of curvature $\mathbb{C}$ (§5.1).

[2] FedAwS and FedUV do not design a specific method to initialize embedding vectors.

Eq. (5) indicates that to minimize the misclassification probability, we should select a modality combination for each user that has the largest product of density $\mathbb{D}$ and ρ (*i.e.,* minimum distance to modality combinations chosen by others in the feature space). However, this is non-trivial under the privacy-preserving FL setup due to the interdependence between selecting a modality combination and determining the value of ρ, which presents a chicken-egg dilemma as the latter depends on the modality selection of other users. The brute force method that considers all possible modality combinations for all users has an exponential time complexity of $O(M^N)$, where M and N are the numbers of modality combinations and users, respectively. For instance, it requires 63^{41} (~6×10^{73}) computations for the `Throw` dataset, which is impractical.

We propose a heuristic algorithm for efficiently selecting the best modality combination for each user. It is motivated by the fact that $\lim_{\mathbb{D}\to\infty} P = 0$. Thus, after density-based modality ordering (§5.1), we allow some users having high-density modality combinations to initialize their class embedding vectors first. Our selection criterion is as follows. All users report the highest density value $\mathbb{D}_{max}$ of their modality combinations to the server. The server then has a set $\{\mathbb{D}_{max}\}$ and can calculate its mean μ and standard deviation σ.

Let Q_{top} be the set of users who directly use their highest-density modality combination to initialize their class embedding vector. Those initial vectors will be reported along with R to the server for calculating ρ for the remaining users. Note that, as introduced above, the reported vectors contain only the mean value of users' input embedding vectors, not the original embedding vectors. Q_{top} is defined as $\left\{i|\mathbb{D}^i_{max} > \mu + k\sigma\right\}$, where k is a parameter that balances the impact of within-user modality ordering and between-user modality selection. The smaller k is, the more users rely solely on density, which may lead to suboptimal performance. On the other hand, the larger k is, the fewer users initialize their class embedding vectors at this point, resulting in potentially inaccurate calculation of ρ for the remaining users. We empirically find that $k = 0.5$ achieves the best balance. We will thoroughly evaluate the impact of k on MetaFL's performance in §6.4.

To select the modality combination for users not in Q_{top}, the server ranks them in descending order based on their highest density $\mathbb{D}_{max}$. Following this order, the server selects the modality combination that minimizes misclassification probability P based on Eq. (5). The minimum distance ρ is calculated with the already initialized class embedding vectors. This design allows users with a lower $\mathbb{D}_{max}$ to have a better chance of obtaining an optimal P for participating in the process later when more users have initialized their class embedding vectors. Our algorithm has a time complexity of $O(N)$, significantly lower than the brute force one.

Combining within-user modality ordering and between-user modality selection can increase the accuracy of FedUV and FedAwS to 94.1% and 91.9% on the `Throw` dataset, respectively, a ~9% improvement compared to utilizing only within-user modality ordering, which is shown in Table 2.

5.3 Personalized Vector Generation

Problem. FL-based models such as FedAwS may still lead to privacy leakage by sharing class embedding vectors with the server [45]. FedUV addresses this issue by utilizing ECCs to increase the distance of class embedding vectors for different users without revealing them to the server, which is orthogonal to our proposed scheme via modality selection. An ideal design would maximize the distance between any pair of ECCs created independently by users. It can in turn maximize the distance between any pair of class embedding vectors, thereby improving authentication performance.

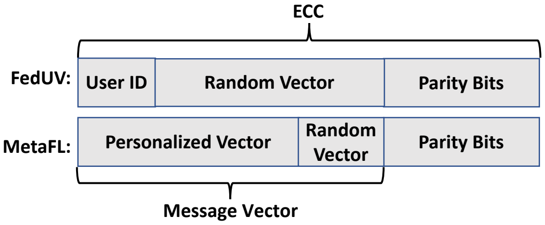


Figure 7: ECC structures of FedUV and MetaFL.

Challenges. To address privacy concerns without sacrificing authentication accuracy, we should carefully design the message vector for each user, which is used to generate an ECC. Although sharing information between users and the server can potentially achieve a long distance between ECCs of different users, it may lead to privacy leakage. On the other hand, letting each user individually create the message vector preserves privacy but may not be able to maximize the distance between ECCs generated by different users. Moreover, increasing the length of generated ECCs may result in better authentication accuracy, but it makes the model complex, increasing communication overhead and inference time. Thus, it is challenging to strike a balance between authentication accuracy, privacy protection, and model complexity.

Our Approach. We leverage the initialized class embedding vector to create a personalized vector as part of the message vector, which is used to generate an ECC. The ECC structures of MetaFL and FedUV are depicted in Figure 7. Different from FedUV, where most part of the message vector is randomly generated, MetaFL introduces a unique personalized vector to increase the distance between ECCs of different users. By utilizing it, MetaFL does not need the user ID as in FedUV, which leads to the minimum Hamming distance between two consecutive IDs being only 1.

Our design is motivated by the fact that when selecting the modality combination for each user to initialize the class embedding vector, we tend to prioritize those with a high ρ (*i.e.,* having a long distance from other class embedding vectors, as defined in §5.2). We convert the initialized class embedding vector w^I to an e-dimension vector using a linear transfer function: $w^I \to \mathbb{R}^e$. Next, we utilize Otsu's method [7] to convert $\mathbb{R}^e$ into a binary vector. In this way, each user may generate a unique personalized vector with a long distance from others. However, we cannot directly use the personalized vector as the message vector as this may cause privacy concerns. Given that the server knows the initial embedding vector of each user (§5.2), it can potentially obtain the personalized vector and its generated ECC by applying the same approach. Thus, the input embedding vector $g_\theta(x)$, which contains the information of raw data x, may be compromised (§7).

To mitigate the privacy concern, we include an r-bit random vector in the message vector, resulting in 2^r possible combinations per user. r is designed to prevent ECCs from being known by the

Models	127 (64)	255 (71)	511 (67)
FedUV	25/1.2	74/2.1	178/2.8
MetaFL*	63/2.5	127/2.7	224/4.2
MetaFL	53/2.1	107/2.4	207/3.8

Table 3: The average/standard deviation Hamming distance of ECCs with different code (message) lengths generated by FedUV, MetaFL*, and MetaFL for all users. MetaFL* refers to MetaFL without adding the random vector.

server, similar to adding noise. We can adjust r to balance the tradeoff between privacy protection and authentication accuracy. For example, a reduced value of r inherently raises concerns regarding privacy breaches. Meanwhile, this results in an increased length of personalized vectors, which leads to greater distance between different ECCs, thereby boosting authentication accuracy. We will thoroughly investigate this tradeoff in §6.4. Finally, each user generates the ECC by concatenating the message vector and the resulting parity bits.

To validate the effectiveness of the personalized vector, we experiment with $r = 20$ to compare the average Hamming distance of ECCs of all users with different code and message lengths generated by FedUV, MetaFL without adding the random vector, and MetaFL. We adopt the same setup as FedUV, setting the ECC code lengths to 127, 255, and 511, with corresponding message lengths of 64, 71, and 67, respectively. We show the results in Table 3. Although the ECCs generated by MetaFL sacrifice some distance to preserve privacy compared to the version without adding the random vector, it still has a larger Hamming distance than FedUV by ~30. As we will show next, MetaFL makes the ECC with a code length of 127 an effective choice for achieving satisfactory authentication accuracy. However, this code length is not sufficient for FedUV.

6 Performance Evaluation

In this section, we extensively evaluate the performance of MetaFL by comparing it with FedAwS [103], FedUV [38], and the central model, as introduced in §2.1.

6.1 Experiment Setup

Implementation. We implement a prototype of MetaFL with PyTorch [78] for model training and the BCH algorithm [12] for generating ECCs. FL involves multiple mobile devices (as many as 41 in our case) collaboratively training a model with the server. Hence, we emulate the FL training process on a machine with an RTX 3080 GPU. This approach is commonly adopted by previous work [47, 50]. Moreover, we measure the inference time on an Nvidia Jetson Xavier NX, whose GPU is comparable [25] to that of the Oculus Quest 2 VR headset, one of the most popular mobile VR headsets [92]. We design a four-layer CNN model, which has been demonstrated as a suitable choice for VR authentication [57, 63], and train it using the SGD optimizer [13]. The output of the CNN is a 128-dimensional embedding vector ($g_\theta(x)$ in §2.1). We randomly select five users in each round with two local training epochs to avoid overfitting [38, 75].

We set the selection parameter k to 0.5, the length of the random vector r to 20, and the length of message/ECC to 64/127 unless specified otherwise and evaluate the impact of different choices for these parameters on authentication performance. We re-implement FedAwS and FedUV as their source code is not publicly available. Our re-implementations align well with their reported results. We use default parameter settings in the original papers to ensure consistency and train them with data from all available modalities.

Datasets	Users	Traces	Samples	Duration (s)
Throw [65]	41	20	135	3
Point [80]	22	520	200	2
Grab [80]	22	520	200	2
Walk [80]	22	160	400	4
Type [80]	22	30	2000	20
Watch [57]	11	3	4800	40

Table 4: Summary of six different datasets.

Datasets. To understand the effectiveness of MetaFL across different authentication tasks and modalities, we evaluate its performance on six public VR authentication datasets, listed in Table 4. A brief overview of each dataset is as follows.

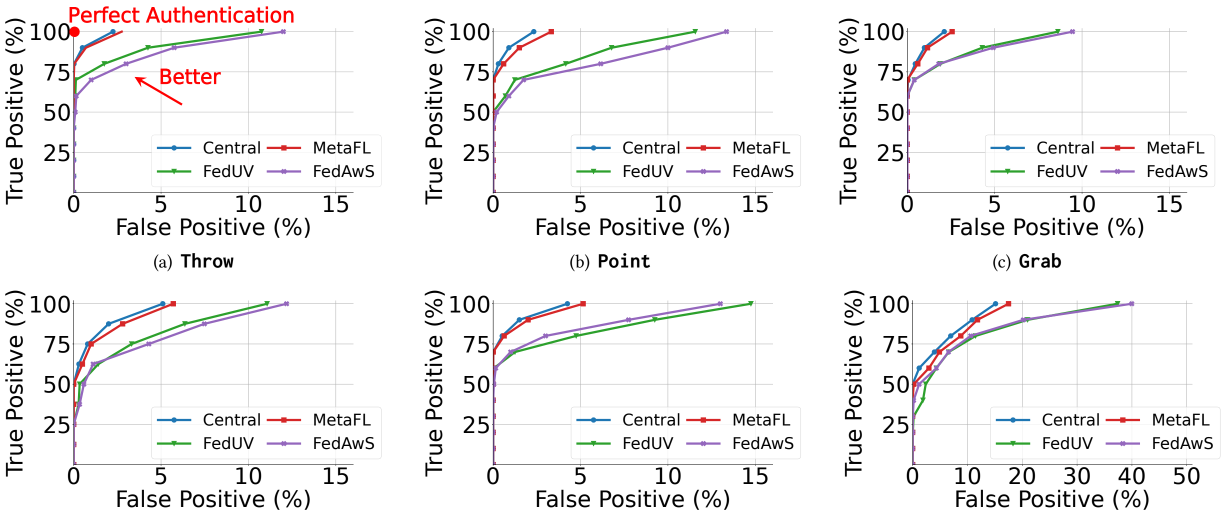
- The `Throw` (Th) dataset [65] (introduced in §2.2) comprises traces collected from 41 users throwing a virtual ball in VR with Oculus Quest 2 on two different days. It contains six modalities: the position and orientation of the headset and both controllers.
- Pfeuffer *et al.* [80] released four datasets collected from 22 users performing different tasks with the HTC VIVE VR headset on two different days. They have eight modalities: gaze positions in the headset and virtual world coordinate systems, as well as the six modalities of the `Throw` dataset. Details of the four datasets are as follows. (1) `Point` (Pt): Pointing to objects with the controller. (2) `Grab` (Gr): Clicking and releasing objects with the controller. (3) `Walk` (Wl): Walking through virtual paths. (4) `Type` (Ty): Typing random sentences on a virtual keyboard.
- The dataset released by Liebers *et al.* [57] consists of 11 users who watched a visual stimulus's movement three times with HTC VIVE, which we refer to as `Watch` (Wt). It contains four modalities: gaze positions in the headset and virtual world coordinate systems, gaze look-at point, and head orientation.

We conduct three-fold cross-validation on `Watch` as it has only three traces and five-fold cross-validation on others.

Evaluation Metrics. Besides the accuracy, we plot the receiver operating characteristic (ROC) curve [43], an important tool for evaluating authentication performance [89]. It shows the true positive rate (TPR) against the false positive rate (FPR) at various discrimination thresholds. Ideally, a perfect system should achieve 100% TPR while maintaining 0% FPR (*i.e.,* the upper left corner of the ROC curve). Therefore, the closer the ROC curve of a model is to that corner, the better its authentication performance. With a ROC curve, we can determine the equal error rate (EER), which represents the point on the curve where the false negative rate (*i.e.,* 1-TPR) equals the FPR. EER provides a single value that jointly considers TPR and FPR. The lower the EER, the better the authentication performance of the model. Additionally, we measure the inference time of each model (*i.e.,* the time from processing a test input to making a decision).

6.2 Authentication Performance

Figure 8 depicts the ROC curves of the four models on six datasets. MetaFL achieves a TPR close to that of the central model under



(a) **Throw** (b) **Point** (c) **Grab**

(d) **Type** (e) **Watch** (f) **Walk**

Figure 8: ROC curves of four models on six datasets. The closer the ROC curve of a model is to the upper left corner, the better its authentication performance will be. Note that the x-axis of (f) has a different scale than others.

Models	Throw				Point				Grab				Type				Watch				Walk			
	A	T @ F		E	A	T @ F		E	A	T @ F		E	A	T @ F		E	A	T @ F		E	A	T @ F		E
		0.5	0.1			0.5	0.1			0.5	0.1			0.5	0.1			0.5	0.1			0.5	0.1	
MetaFL	98	80	72	1.4	97	71	68	1.6	97	71	70	1.3	92	59	50	2.9	95	72	70	2.6	73	58	55	9.3
FedUV	76	56	51	7.1	74	51	50	8.4	76	61	59	7.1	71	25	23	9.4	70	52	42	9.7	53	25	23	15
FedAwS	74	46	41	7.9	69	43	41	10	75	61	59	7.5	70	26	25	9.9	72	55	45	8.9	56	31	29	14
Central	98	81	74	1.1	98	72	70	1.2	97	71	70	1.1	93	61	50	2.6	95	72	70	2.1	75	60	55	8.9

Table 5: Accuracy (A), TPR (T) under 0.5 or 0.1 target FPR (F), and EER (E) in % of four models on six datasets.

different FPRs on all datasets. MetaFL outperforms FedUV and FedAwS, and their gap of FPR grows with an increasing TPR. For example, MetaFL can achieve 100% TPR with only 2–5% FPR on all datasets except for **Walk**, while FedUV and FedAwS lead to 10–15% FPR to achieve the same. For the **Walk** dataset, even the central model performs poorly (*e.g.,* ~15% FPR when TPR is 100%). A possible reason is that the movements of hands, head, and gaze when users walk may not present distinctly unique patterns [80]. In this scenario, utilizing gait may be a better choice [84].

The discrimination threshold is often determined by the target FPR of an authentication system, which is usually low (*e.g.,* <1% [89]) because the cost of FPR is significant in practice [42, 43]. Thus, we report the TPR for a target FPR of 0.5% and 0.1%, respectively, and the accuracy and EER of different models on six datasets in Table 5. MetaFL has a comparable accuracy and EER with the central model, with a small gap of <2% for accuracy and <0.5% for EER, and outperforms FedUV and FedAwS by improving 17–28% in accuracy and reducing 4.7–8.4% in EER. Moreover, for a low target FPR, the TPR of MetaFL remains close to that of the central model with a small gap of <2%, while showing an improvement of 11–34% compared to FedUV and FedAwS.

Figure 9 shows the inference time of four models on six datasets. MetaFL can conduct an authentication in <250ms, which is comparable to FedUV, and incurs only ~60ms extra latency compared to the central model. The inference time on the **Type** dataset is higher than others, because it has more modalities than **Throw** and **Watch** and more data samples for each motion than **Point**, **Grab**, and **Walk**.

6.3 Scalability Analysis

We next evaluate the scalability of MetaFL, by conducting experiments on the **Throw** dataset, which has the largest number of users among the six datasets. Figure 10 shows the accuracy[3] of four models with different numbers of users. As the number of users increases, the accuracy of MetaFL maintains a small gap with the central model. When there are 41 users, MetaFL still achieves an accuracy of >98%, whereas the accuracy of FedUV and FedAwS drops to below 80%.

To evaluate the scalability of MetaFL from another perspective, we scale down the number of modalities. We randomly select 3–5 modalities for authentication on each of the datasets (excluding **Walking** since it has only 4 modalities) and evaluate the accuracy of MetaFL. As shown in Figure 11, even when only three modalities are available, the average accuracy of MetaFL decreases by only <5%, compared to using all modalities. Moreover, the maximum accuracy remains consistent in both scenarios, mainly because more than 80% of users select only one or two modalities.

[3] We verify that high accuracy indicates low EER for all experiments.

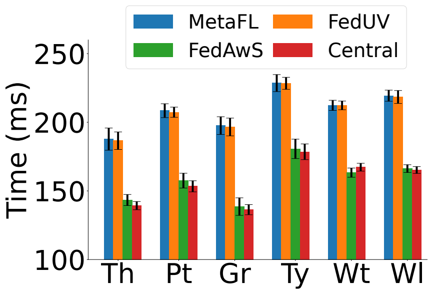


Figure 9: Inference time of four models on six datasets. The test device is Nvidia Jetson Xavier NX.

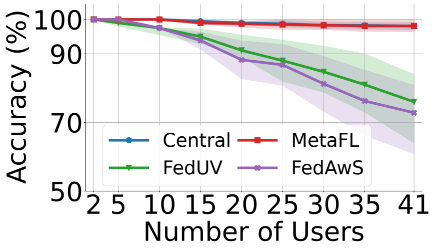


Figure 10: Four models' accuracy with different numbers of users on the `Throw` dataset (The band shows 95% CIs).

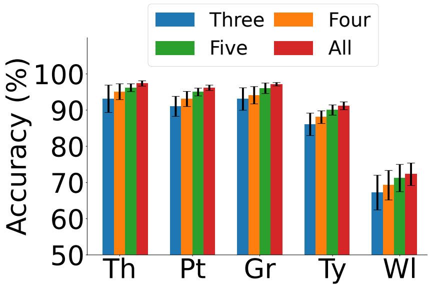


Figure 11: Accuracy of MetaFL on five datasets when only three, four, five, or all modalities are available.

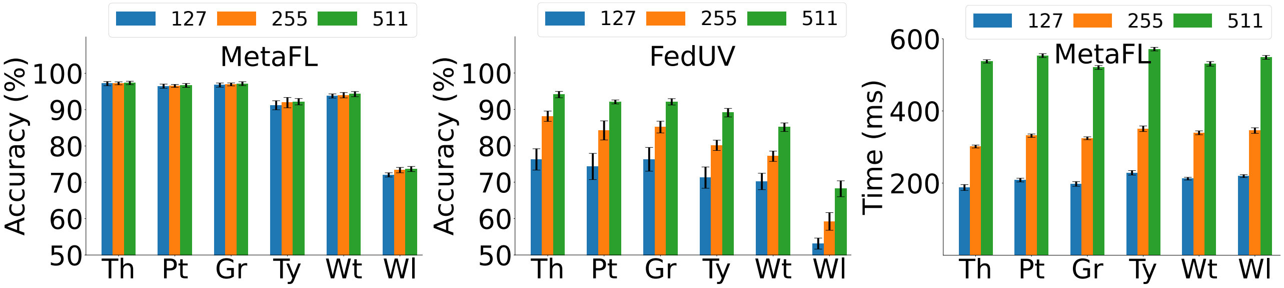


Figure 12: Accuracy of MetaFL and FedUV, and inference time of MetaFL with different lengths of ECC on six datasets.

6.4 Robustness and Sensitivity Analysis

We then analyze the robustness and sensitivity of MetaFL for the following parameters.

Length of ECC affects both accuracy and inference time of ECC-based authentication models. Figure 12 shows the accuracy of MetaFL and FedUV and the inference time of MetaFL with different ECC lengths on six datasets. The inference time of FedUV is similar to MetaFL. Benefiting from judiciously selecting modality combinations and generating personalized vectors for users, 127-bit ECCs are enough for MetaFL to achieve satisfactory performance with only a <2% gap in accuracy to that of 511-bit ECCs, and reduce the inference time by 300–400 ms compared to 511-bit ECCs. In contrast, the accuracy of FedUV with 511-bit ECCs is still lower than that of MetaFL with 127-bit ECCs (*e.g.,* 85.0% for FedUV *vs* 94.2% for MetaFL on the `Watch` dataset).

Selection Parameter k balances the impact of within-user modality ordering and between-user modality selection (§5.2). As demonstrated in Figure 13, setting k to 0.5 leads to better performance across six datasets. Moreover, the value of k has a limited impact on the datasets that achieve high authentication accuracy (*e.g.,* `Throw`, `Point`, and `Grab`). However, for datasets with low accuracy (*e.g.,* `Walk`), a small k will significantly decrease authentication accuracy, because their motion data would typically present more randomness (*i.e.,* low density).

Length of Random Vector r**.** As r decreases, MetaFL's accuracy increases at the cost of possible privacy leakage (§5.3). Our evaluation indicates that $r = 20$ strikes a balance between privacy and accuracy with <2% reduction in accuracy compared to $r = 0$ while effectively protecting users' ECCs by creating $2^{20} = 1,048,576$ different message vectors.

6.5 Communication Overhead

We next evaluate the communication overhead of MetaFL using the `Throw` dataset, which has the largest number of users in our setup. The average bandwidth consumption of MetaFL is 4.81 Mbps (SD = 0.4), with ~0.1 Mbps per user on average for 41 users, sufficiently low to facilitate FL model training under current WiFi and 4G LTE networks [74, 75].

Compared with FedUV, MetaFL requires a shorter length of ECC (§6.4), reducing the size of transmitted W (§2.1). For instance, MetaFL with 127-bit ECCs achieves a 76.2% reduction in bandwidth consumption compared to FedUV with 511-bit ECCs, which is 20.23 Mbps on average (SD = 2.7), while maintaining higher accuracy and reducing latency by >300 ms (§6.4).

Different from FedAwS, the bandwidth consumption of MetaFL may not increase with the number of users, as the transmitted W depends solely on the ECC length and dimensions of input embedding vectors (§2.1). Conversely, FedAWS has scalability issues as each client must update the class embedding vector to the server. For example, the bandwidth required by FedAWS escalates from 4.93 Mbps (SD = 0.51) with 25 users to 7.21 Mbps (SD = 0.87) with 41 users, and will continue increasing as the number of users grows.

6.6 Temporal Effect of Modalities

We then investigate the temporal effect of modalities on the authentication performance of MetaFL. Except for `Watch`, all other five datasets are collected on different days (§6.1). Thus, instead of five-fold cross-validation, we use data collected on different days for training and testing.

The results are shown in Figure 14. Under this setup, the accuracy of MetaFL is only <1% lower than five-fold cross-validation,

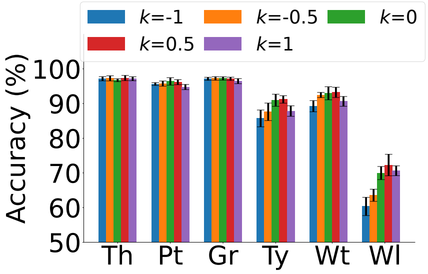


Figure 13: Accuracy of MetaFL with different values of selection parameter k on six different datasets.

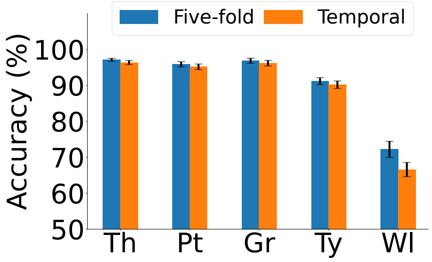


Figure 14: MetaFL's accuracy with five-fold cross-validation and using data on different days for training and testing.

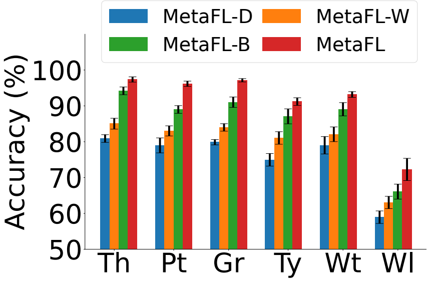


Figure 15: Accuracy of three breakdown versions of MetaFL and full-fledged MetaFL on six different datasets.

except for the `Walk` dataset. The reduction is because this setting leads to less training but more testing samples than five-fold cross-validation, which causes the accuracy of FedUV and FedAwS to drop by >8% (figure not shown due to space limit). On the `Walk` dataset, the accuracy of MetaFL decreases by ~5% compared to five-fold cross-validation. The reduction may be again attributed to the collected modalities in the walking tasks are not suitable for authentication (§6.2). It has a more significant impact on the accuracy of FedUV and FedAwS with a >20% decrease. In summary, the above results demonstrate that the modality combination selected by MetaFL is robust and not affected by the possible temporal changes across different days.

6.7 Component-wise Analysis

We finally evaluate the effectiveness of each key component in MetaFL by implementing three breakdown versions of MetaFL and comparing them with the fully-fledged MetaFL.

- MetaFL-D: Users select modality combinations with the highest density after denoising (*i.e.,* without considering the density both before and after denoising).
- MetaFL-W: Users select modality combinations with the highest density using only within-user modality ordering.
- MetaFL-B: Users select modality combinations based on MetaFL without generating the personalized vector.

Figure 15 shows the accuracy of these four versions of MetaFL on six datasets. By considering the density both before and after denoising, MetaFL can improve the accuracy by ~4%. Adding between-user modality selection increases accuracy by 4–8%. Finally, personalized vector generation can further enhance accuracy by 4–9%.

7 Security and Privacy Analysis

In this section, we conduct the security and privacy analysis of MetaFL against several attacks described in §3. Specifically, security concerns arise from the attackers gaining unauthorized access, which is client-based attacks. Meanwhile, privacy concerns stem from the exposure of users' biometric data, which is the target of server-based and network-based attacks.

Impersonation Attacks. Taking into account real-world scenarios, such attackers may not necessarily be enrolled in the system. Therefore, we assess the capability of MetaFL to detect attackers whose data have not been incorporated into the training set. For all six datasets, we randomly select 50% of users for training, and the remaining users attempt to authenticate on each trained model. Our results reveal that the accuracy of MetaFL remains >99% across all settings, indicating its robust defense against such attacks.

Mimicry Attacks. Leveraging multi-modal biometric data, MetaFL provides resilient protection against such attacks for two reasons: (1) simultaneously replicating the movements of multiple body parts is difficult [87] and (2) gaze motion (if applicable) is resistant to observation as the user's face is covered by the headset [62, 108].

To verify if MetaFL can defend against such attacks, we conducted an IRB-approved user study with 20 users (7 females, 13 males). We developed a VR game in which users can throw an axe into a dartboard with the HTC Vive Pro Eye headset. We collect seven modalities: gaze positions and the same six modalities of the `Throw` dataset. During experiments, participants were paired. Each user performed the enrollment motion (1.5 seconds) 40 times while being observed by their partner. Each participant then attempted to mimic their partner's motion 20 times. The enrollment motions were used to train the model, and the mimic motions were used for conducting mimicry attacks. Considering some headsets may not offer gaze tracking, we conducted the training/testing in two scenarios: one with gaze motion and the other without. Our results reveal that MetaFL's accuracy maintains 100% in both scenarios, demonstrating its capacity to defend against such attacks.

Model Inversion Attacks. In MetaFL, each user updates the projection matrix W utilizing its ECC v_y and input embedding vector $g_\theta(x)$ during the training phase, similar to FedUV (§2.1). The $g_\theta(x)$ is the sensitive information we should preserve since the server can conduct model inversion attacks to obtain the information about the raw data x. Although $g_\theta(x)$ is always kept locally, if the server gains knowledge of ECC v_y, it can potentially infer $g_\theta(x)$ by conducting the following two attacks.

#1: By executing the gradient-based inversion attack, the server could potentially obtain the relationship between v_y and $g_\theta(x)$ of the user during training, thus deducing $g_\theta(x)$.

#2: Recall that a local loss of zero implies $Wg_\theta(x) = v_y$ (§2.1). Given this relationship, the server could calculate the value of $g_\theta(x)$. Even though the server may not acquire the exact value of the local loss, it could make a reasonable assumption that the client's local loss in the last communication round is zero to infer $g_\theta(x)$.

The above analysis reflects that to protect $g_\theta(x)$, the key is to prevent the server from discerning the client's ECC v_y. To achieve this, we introduce a random vector into the ECC design (§5.3). Specifically, we integrate a 20-bit random vector in MetaFL, leading to 1,048,576 different combinations, significantly minimizing the server's probability of identifying the true ECC.

Network-based Attacks. As analyzed above, since the data transmitted between clients and the server is the projection matrix W rather than raw behavioral biometrics, attackers cannot directly infer information about the raw data. Even if they can obtain some information about W, similar to model inversion attacks, without knowing the client's ECC, they may still be unable to deduce information from the raw biometrics.

8 Discussion

Deployment. In this paper, we highlight the potential of FL for authenticating VR users with dynamic modalities, particularly behavioral biometrics, which enhance accessibility and robustness over static methods. Here, we discuss a few practical considerations for deploying MetaFL.

Handling Users Joining and Departure: In real-world scenarios, new users joining and existing users leaving the system is common. When new users join, as the trained model has gained the ability to distinguish existing users, MetaFL does not need to re-train from scratch. Instead, it can fine-tune the existing model to accommodate new users. When users leave the system, to completely eliminate their training records, we can employ emerging federated unlearning techniques [37].

Integrating Traditional Authentication Methods: While behavioral biometrics offer a resilient and convenient mode of authentication, it might not succeed in all scenarios. Hence, we could resort to traditional authentication methods such as passwords as a backup, similar to what Apple's Face ID does [5]. This complimentary setup allows users to fall back to their password if behavioral biometric-based authentication fails, providing an additional security layer and preventing user lockout.

Addressing Biometric Variability: Given the dynamic nature of behavioral biometrics, they may alter over time due to various factors. For instance, an injury such as a sprained wrist could change an individual's pattern of hand movements. To account for these changes, we could allow users to update their enrollment information, mirroring the concept of password resets.

Defending Against Other Attacks. MetaFL currently focuses on addressing the challenges of accurately authenticating VR users through FL. Thus, it prioritizes prevalent threats in behavioral biometric-based VR authentication (§3). To explore strategies for enhancing MetaFL's defenses against an expanded array of attacks, we consider a scenario where attackers deploy malicious software on users' headsets to capture (part of) data during enrollment. Such tactics could enable them to execute replay or synthesis attacks [16]. For these attacks, we can design a dynamic authentication scheme wherein the position of targets (*e.g.,* virtual ball) during authentication is altered each time, similar to the dynamic virtual keyboard layouts [1] during password entry.

Defending Against Privacy Leakage. MetaFL employs ECCs to safeguard users from sharing class embedding vectors with the server, offering enhanced privacy protection compared to traditional FL approaches such as FedAvg [64]. However, MetaFL still requires users to upload the highest density value of their modality combinations and the mean of their input embedding vectors. Despite this, the impact of privacy leakage is constrained since the server does not have access to the raw data or embedding vectors and remains unaware of which modality combination the clients select. While the server's ability to infer modality information from these values is currently under-explored, we anticipate its risk to be low. A promising direction for future research is to study the incorporation of other privacy-preserving techniques into MetaFL to achieve an optimal balance between privacy and utility [102].

9 Related Work

Security & Privacy in XR. Security and privacy have been extensively studied for XR [17, 18, 39, 85, 99, 101, 105]. Compared to other mobile devices, users wearing headsets are more vulnerable to attacks that, for example, employ facial dynamics [85] or perceptual manipulation [17]. However, few studies have addressed the privacy issues of authentication in XR, which is the key problem that MetaFL aims to solve.

Biometric-based VR User Authentication. There are two types of biometric data for VR authentication: (1) physical biometrics, such as skull [83, 95], muscle structure [16], brain waves [52, 58], speech signal [104], and ear canal [26]; and (2) behavioral biometrics, such as head [56, 65, 80, 106], hand [65, 80], and gaze [56, 106] movements. Existing efforts use centralized training approaches, leading to privacy leakage. In contrast, we leverage FL for privacy-preserving VR authentication.

Research of Federated Learning. Given its ability to preserve privacy while training deep learning models [68, 71, 73], FL has raised the interests of the community with numerous efforts, such as addressing the heterogeneity issue of data [20, 50, 86] and computational recourse [48, 49, 90] on clients, as well as applying it to human activity recognition [74, 75, 93]. In this paper, we benefit from FL to avoid privacy leakage when authenticating VR users.

10 Conclusion

This paper presented the design, implementation, and evaluation of MetaFL, a privacy-preserving authentication framework for VR users that leverages FL on their multi-modal behavioral biometrics. The design of MetaFL is motivated by a fundamental challenge of privacy-preserving user authentication tasks: users have access to only their own data, causing the impact of different modalities to significantly vary on authentication accuracy for different users. To address this, MetaFL intelligently selects modalities for users aiming to minimize the misclassification probability and designs a personalized scheme to further increase authentication accuracy. Our extensive performance evaluation of MetaFL shows that it can drastically improve authentication performance compared to state-of-the-art solutions.

Acknowledgment

We thank the anonymous reviewers and our shepherd for their insightful feedback, as well as the participants of our experiments for their valuable contributions. This work was partially supported by the National Science Foundation under Grant CNS-2235049 and a research award from Cisco.

References

[1] M. Agarwal, M. Mehra, R. Pawar, and D. Shah. Secure Authentication Using Dynamic Virtual Keyboard Layout. In *Proceedings of International Conference & Workshop on Emerging Trends in Technology*, 2011.

[2] D. Aggarwal, J. Zhou, and A. K. Jain. FedFace: Collaborative Learning of Face Recognition Model. In *Proceedings of IEEE International Joint Conference on Biometrics (IJCB)*, 2021.

[3] A. Al Arafat, Z. Guo, and A. Awad. VR-Spy: A Side-Channel Attack on Virtual Key-Logging in VR Headsets. In *IEEE Virtual Reality and 3D User Interfaces (VR)*, 2021.

[4] A. Alzubaidi and J. Kalita. Authentication of Smartphone Users Using Behavioral Biometrics. *IEEE Communications Surveys & Tutorials*, 18(3):1998–2026, 2016.

[5] Apple. About Face ID advanced technology. https://support.apple.com/en-us/HT208108, 2024. [accessed on 10/07/2024].

[6] K. O. Bailey, J. S. Okolica, and G. L. Peterson. User Identification and Authentication Using Multi-modal Behavioral Biometrics. *Computers & Security*, 43:77–89, 2014.

[7] S. L. Bangare, A. Dubal, P. S. Bangare, and S. Patil. Reviewing Otsu's Method for Image Thresholding. *International Journal of Applied Engineering Research*, 10(9):21777–21783, 2015.

[8] M. Barni, G. Droandi, R. Lazzeretti, and T. Pignata. SEMBA: Secure Multi-biometric Authentication. *IET Biometrics*, 8(6):411–421, 2019.

[9] V. N. Boddeti. Secure Face Matching Using Fully Homomorphic Encryption. In *Proceedings of IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS)*, 2018.

[10] P. Bojanowski and A. Joulin. Unsupervised Learning by Predicting Noise. In *International Conference on Machine Learning (ICML)*, 2017.

[11] J. Bonneau, C. Herley, P. C. Van Oorschot, and F. Stajano. The Quest to Replace Passwords: A Framework for Comparative Evaluation of Web Authentication Schemes. In *Proceedings of IEEE Symposium on Security and Privacy (S&P)*, 2012.

[12] R. C. Bose and D. K. Ray-Chaudhuri. On a Class of Error Correcting Binary Group Codes. *Information and Control*, 3(1):68–79, 1960.

[13] L. Bottou. Large-scale Machine Learning with Stochastic Gradient Descent. In *Proceedings of International Conference on Computational Statistics*, 2010.

[14] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. LOF: Identifying Density-based Local Tutliers. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 93–104, 2000.

[15] P. R. Center. How Americans View Data Privacy. https://www.pewresearch.org/internet/2023/10/18/how-americans-view-data-privacy/, 2023. [accessed on 10/07/2024].

[16] Y. Chen, Z. Yang, R. Abbou, P. Lopes, B. Y. Zhao, and H. Zheng. User Authentication via Electrical Muscle Stimulation. In *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI)*, 2021.

[17] K. Cheng, J. F. Tian, T. Kohno, and F. Roesner. Exploring User Reactions and Mental Models Towards Perceptual Manipulation Attacks in Mixed Reality. In *Proceedings of the USENIX Security Symposium*, 2023.

[18] R. Cheng, S. Chen, and B. Han. Toward Zero-trust Security for the Metaverse. *IEEE Communications Magazine*, 62(2):156–162, 2023.

[19] D. Comaniciu and P. Meer. Mean Shift Analysis and Applications. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1999.

[20] Y. Deng, W. Chen, J. Ren, F. Lyu, Y. Liu, Y. Liu, and Y. Zhang. TailorFL: Dual-Personalized Federated Learning under System and Data Heterogeneity. In *Proceedings of ACM SenSys*, 2022.

[21] D. K. Dennis, T. Li, and V. Smith. Heterogeneity for the Win: One-Shot Federated Clustering. In *Proceddings of ICML*, 2021.

[22] H. Duan, J. Li, S. Fan, Z. Lin, X. Wu, and W. Cai. Metaverse for Social Good: A University Campus Prototype. In *Proceedings of ACM International Conference on Multimedia (MM)*, 2021.

[23] A. Ferlini, D. Ma, R. Harle, and C. Mascolo. EarGate: Gait-based User Identification with In-ear Microphones. In *Proceedings of ACM MobiCom*, 2021.

[24] M. Fredrikson, S. Jha, and T. Ristenpart. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In *Proceedings of ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2015.

[25] GadgetVersus. Qualcomm Adreno 650 vs Nvidia Jetson Xavier NX GPU. https://bit.ly/42srjGz, 2023. [accessed on 10/07/2024].

[26] Y. Gao, W. Wang, V. V. Phoha, W. Sun, and Z. Jin. EarEcho: Using Ear Canal Echo for Wearable Authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):1–24, 2019.

[27] Y. A. Ghassabeh. A Sufficient Condition for the Convergence of the Mean Shift Algorithm with Gaussian Kernel. *Journal of Multivariate Analysis*, 135:1–10, 2015.

[28] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran. An Efficient Framework for Clustered Federated Learning. In *Proceddings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[29] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran. An Efficient Framework for Clustered Federated Learning. *IEEE Transactions on Information Theory*, 68(12):8076–8091, 2022.

[30] M. Golla, G. Ho, M. Lohmus, M. Pulluri, and E. M. Redmiles. Driving 2FA Adoption at Scale: Optimizing Two-Factor Authentication Notification Design Patterns. In *Proceedings of the USENIX Security Symposium*, 2021.

[31] Google. Learn how Google improves speech models. https://support.google.com/assistant/answer/11140942?hl=en#zippy=%2Cfederated-learning, 2024. [accessed on 10/07/2024].

[32] S. D. Gordon, D. Starin, and A. Yerukhimovich. The More The Merrier: Reducing the Cost of Large Scale MPC. In *Proceedings of Annual International Conference on the Theory and Applications of Cryptographic Techniques (Eurocrypt)*, 2021.

[33] N. Görnitz, A. Porbadnigk, A. Binder, C. Sannelli, M. Braun, K.-R. Müller, and M. Kloft. Learning and Evaluation in Presence of Non-iid Label Noise. In *Artificial Intelligence and Statistics (AISTATS)*, 2014.

[34] F. Granqvist, M. Seigel, R. Van Dalen, A. Cahill, S. Shum, and M. Paulik. Improving On-Device Speaker Verification Using Federated Learning with Privacy. In *Proceddings of Interspeech*, 2020.

[35] H. Habib, P. E. Naeini, S. Devlin, M. Oates, C. Swoopes, L. Bauer, N. Christin, and L. F. Cranor. User Behaviors and Attitudes Under Password Expiration Policies. In *Proceedings of Symposium on Usable Privacy and Security (SOUPS)*, 2018.

[36] J. F. Hair Jr, W. C. Black, B. J. Babin, and R. E. Anderson. Multivariate Data Analysis. 2010.

[37] A. Halimi, S. R. Kadhe, A. Rawat, and N. B. Angel. Federated Unlearning: How to Efficiently Erase a Client in FL? In *Proceddings of ICML*, 2022.

[38] H. Hosseini, H. Park, S. Yun, C. Louizos, J. Soriaga, and M. Welling. Federated Learning of User Verification Models Without Sharing Embeddings. In *Proceedings of ICML*, 2021.

[39] J. Hu, A. Iosifescu, and R. LiKamWa. Lenscap: Split-process Framework for Fine-Grained Visual Privacy Control for Augmented Reality Apps. In *Proceedings of ACM MobiSys*, 2021.

[40] Y. Huang, S. Gupta, Z. Song, K. Li, and S. Arora. Evaluating Gradient Inversion Attacks and Defenses in Federated Learning. In *Proceedings of Conference on Neural Information Processing Systems (NIPS)*, 2021.

[41] A. Jain, R. Bolle, and S. Pankanti. *Introduction to Biometrics*. Springer, 1996.

[42] A. K. Jain and K. Nandakumar. Biometric Authentication: System Security and User Privacy. *Computer*, 45(11):87–92, 2012.

[43] A. K. Jain, A. Ross, and S. Prabhakar. An Introduction to Biometric Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):4–20, 2004.

[44] I. T. Jolliffe and J. Cadima. Principal Component Analysis: a Review and Recent Developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[45] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al. Advances and Open Problems in Federated Learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[46] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*, 60(6):84–90, 2017.

[47] F. Lai, X. Zhu, H. V. Madhyastha, and M. Chowdhury. Oort: Efficient Federated Learning via Guided Participant Selection. In *Proceedings of USENIX OSDI*, 2021.

[48] A. Li, J. Sun, P. Li, Y. Pu, H. Li, and Y. Chen. Hermes: an Efficient Federated Learning Framework for Heterogeneous Mobile Clients. In *Proceedings of ACM MobiCom*, 2021.

[49] A. Li, J. Sun, X. Zeng, M. Zhang, H. Li, and Y. Chen. Fedmask: Joint Computation and Communication-efficient Personalized Federated Learning via heterogeneous Masking. In *Proceedings of ACM Sensys*, 2021.

[50] C. Li, X. Zeng, M. Zhang, and Z. Cao. PyramidFL: Fine-grained Data and System Heterogeneity-aware Client Selection for Efficient Federated Learning. In *Proceedings of ACM MobiCom*, 2022.

[51] S. Li, A. Ashok, Y. Zhang, C. Xu, J. Lindqvist, and M. Gruteser. Whose Move is it Anyway? Authenticating Smart Wearable Devices Using Unique Head Movement Patterns. In *Proceedings of IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2016.

[52] S. Li, S. Savaliya, L. Marino, A. M. Leider, and C. C. Tappert. Brain Signal Authentication for Human-Computer Interaction in Virtual Reality. In *IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, 2019.

[53] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith. Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

[54] X. Li, R. Dowsley, and M. De Cock. Privacy-preserving Feature Selection with Secure Multiparty Computation. In *Proceedings of ICML*, 2021.

[55] Y. Li and Y. Yuan. Convergence Analysis of Two-layer Neural Networks with Relu Activation. In *Proceedings of Conference on Neural Information Processing Systems (NIPS)*, 2017.

[56] J. Liebers, M. Abdelaziz, L. Mecke, A. Saad, J. Auda, U. Gruenefeld, F. Alt, and S. Schneegass. Understanding User Identification in Virtual Reality Through Behavioral Biometrics and the Effect of Body Normalization. In *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI)*, 2021.

[57] 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 ACM Symposium on Virtual Reality Software and Technology (VRST)*, 2021.

[58] F. Lin, K. W. Cho, C. Song, W. Xu, and Z. Jin. Brain Password: A Secure and Truly Cancelable Brain Biometrics for Smart Headwear. In *Proceedings of ACM MobiSys*, 2018.

[59] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep Hypersphere Embedding for Face Recognition. In *Proceedings of IEEE/CVF CVPR*, 2017.

[60] W. Liu, Y. Wen, Z. Yu, and M. Yang. Large-margin Softmax Loss for Convolutional Neural Networks. In *Proceedings of ICML*, 2016.

[61] M. Luo, F. Chen, D. Hu, Y. Zhang, J. Liang, and J. Feng. No Fear of Heterogeneity: Classifier Calibration for Federated Learning with Non-iid Data. In *Proceedings of Conference on Neural Information Processing Systems (NIPS)*, 2021.

[62] S. Luo, A. Nguyen, C. Song, F. Lin, W. Xu, and Z. Yan. OcuLock: Exploring Human Visual System for Authentication in Virtual Reality Head-mounted Display. In *Proceedings of Network and Distributed System Security Symposium (NDSS)*, 2020.

[63] F. Mathis, H. I. Fawaz, and M. Khamis. Knowledge-driven Biometric Authentication in Virtual Reality. In *Proceedings of ACM CHI*, 2020.

[64] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

[65] R. Miller, N. K. Banerjee, and S. Banerjee. Using Siamese Neural Networks to Perform Cross-System Behavioral Authentication in Virtual Reality. In *Proceedings of IEEE Virtual Reality and 3D User Interfaces (VR)*, 2021.

[66] R. Miller, N. K. Banerjee, and S. Banerjee. Combining Real-World Constraints on User Behavior with Deep Neural Networks for Virtual Reality (VR) Biometrics. In *Proceedings of IEEE VR*, 2022.

[67] R. Miller, N. K. Banerjee, and S. Banerjee. Temporal Effects in Motion Behavior for Virtual Reality (VR) Biometrics. In *Proceedings of IEEE VR*, 2022.

[68] F. Mo, H. Haddadi, K. Katevas, E. Marin, D. Perino, and N. Kourtellis. PPFL: Privacy-preserving Federated Learning with Trusted Execution Environments. In *Proceedings of ACM Mobisys*, 2021.

[69] G. Munilla Garrido, V. Nair, and D. Song. SoK: Data Privacy in Virtual Reality. https://arxiv.org/pdf/2301.05940.pdf, 2023. [accessed on 10/07/2024].

[70] T. Mustafa, R. Matovu, A. Serwadda, and N. Muirhead. Unsure How to Authenticate on Your VR Headset? Come on, Use Your Head! In *ACM International Workshop on Security and Privacy Analytics*, 2018.

[71] C. Niu, F. Wu, S. Tang, L. Hua, R. Jia, C. Lv, Z. Wu, and G. Chen. Billion-Scale Federated Learning on Mobile Clients: A Submodel Design with Tunable Privacy. In *Proceedings of ACM MobiCom*, 2020.

[72] A. Orvieto, J. Kohler, D. Pavllo, T. Hofmann, and A. Lucchi. Vanishing Curvature and the Power of Adaptive Methods in Randomly Initialized Deep Networks. https://arxiv.org/pdf/2106.03763.pdf, 2021.

[73] X. Ouyang, X. Shuai, Y. Li, L. Pan, X. Zhang, H. Fu, X. Wang, S. Cao, J. Xin, H. Mok, et al. ADMarker: A Multi-Modal Federated Learning System for Monitoring Digital Biomarkers of Alzheimer's Disease. In *Proceddings of ACM MobiCom*, 2024.

[74] X. Ouyang, Z. Xie, H. Fu, S. Cheng, L. Pan, N. Ling, G. Xing, J. Zhou, and J. Huang. Harmony: Heterogeneous Multi-Modal Federated Learning through Disentangled Model Training. In *Proceedings of ACM MobiCom*, 2023.

[75] X. Ouyang, Z. Xie, J. Zhou, J. Huang, and G. Xing. ClusterFL: A Similarity-Aware Federated Learning System for Human Activity Recognition. In *Proceedings of ACM Mobisys*, 2021.

[76] J. Pareek and J. Jacob. Data Compression and Visualization Using PCA and T-SNE. In *Proceedings of Advances in Information Communication Technology and Computing*, 2021.

[77] H. Park, H. Hosseini, and S. Yun. Federated Learning with Metric Loss. In *Proceedings of International Workshop on Federated Learning for User Privacy and Data Confidentiality*, 2021.

[78] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An Imperative Style, High-performance Deep Learning Library. In *Proceedings of Conference on Neural Information Processing Systems (NIPS)*, 2019.

[79] V. M. Patel, N. K. Ratha, and R. Chellappa. Cancelable Biometrics: A Review. *IEEE Signal Processing Magazine*, 32(5):54–65, 2015.

[80] K. Pfeuffer, M. J. Geiger, S. Prange, L. Mecke, D. Buschek, and F. Alt. Behavioural Biometrics in VR: Identifying People from Body Motion and Relations in Virtual Reality. In *Proceedings of ACM CHI*, 2019.

[81] Qualcomm. Enabling on-device learning at scale. https://www.qualcomm.com/news/onq/2021/10/enabling-device-learning-scale, 2021. [accessed on 10/07/2024].

[82] T. Sattler, B. Leibe, and L. Kobbelt. Fast Image-based Localization Using Direct 2d-to-3d Matching. In *Proceedings of IEEE/CVF ICCV*, 2011.

[83] S. Schneegass, Y. Oualil, and A. Bulling. SkullConduct: Biometric User Identification on Eyewear Computers Using Bone Conduction Through the Skull. In *Proceedings of ACM CHI*, 2016.

[84] Y. Shen, H. Wen, C. Luo, W. Xu, T. Zhang, W. Hu, and D. Rus. GaitLock: Protect Virtual and Augmented Reality Headsets Using Gait. *IEEE Transactions on Dependable and Secure Computing*, 16(3):484–497, 2018.

[85] C. Shi, X. Xu, T. Zhang, P. Walker, Y. Wu, J. Liu, N. Saxena, Y. Chen, and J. Yu. Face-Mic: Inferring Live Speech and Speaker Identity via Subtle Facial Dynamics Captured by AR/VR Motion Sensors. In *Proceedings of ACM MobiCom*, 2021.

[86] J. Shin, Y. Li, Y. Liu, and S.-J. Lee. FedBalancer: Data and Pace Control for Efficient Federated Learning on Heterogeneous Clients. In *Proceedings of ACM MobiSys*, 2022.

[87] S. Stephenson, B. Pal, S. Fan, E. Fernandes, Y. Zhao, and R. Chatterjee. SoK: Authentication in Augmented and Virtual Reality. In *Proceedings of IEEE Symposium on Security and Privacy (S&P)*, 2022.

[88] Z. Su, K. Cai, R. Beeler, L. Dresel, A. Garcia, I. Grishchenko, Y. Tian, C. Kruegel, and G. Vigna. Remote Keylogging Attacks in Multi-user VR Applications. In *Proceedings of USENIX Security Symposium*, 2024.

[89] S. Sugrim, C. Liu, M. McLean, and J. Lindqvist. Robust Performance Metrics for Authentication Systems. In *Proceddings of Network and Distributed Systems Security Symposium (NDSS)*, 2019.

[90] J. Sun, A. Li, L. Duan, S. Alam, X. Deng, X. Guo, H. Wang, M. Gorlatova, M. Zhang, H. Li, et al. FedSEA: A Semi-Asynchronous Federated Learning Framework for Extremely Heterogeneous Devices. In *Proceedings of ACM Sensys*, 2022.

[91] K. Sundararajan and D. L. Woodard. Deep Learning for Biometrics: A Survey. *ACM Computing Surveys (CSUR)*, 51(3):1–34, 2018.

[92] Tech Times. Meta's Oculus Quest 2 is the Top Selling VR Headset of 2021. https://bit.ly/3Ltruuq, 2022. [accessed on 10/07/2024].

[93] L. Tu, X. Ouyang, J. Zhou, Y. He, and G. Xing. FedDL: Federated Learning via Dynamic Layer Sharing for Human Activity Recognition. In *Proceedings of ACM SenSys*, 2021.

[94] L. Van der Maaten and G. Hinton. Visualizing Data Using t-SNE. *Journal of machine learning research*, 9(11), 2008.

[95] R. Wang, L. Huang, and C. Wang. Low-effort VR Headset User Authentication Using Head-reverberated Sounds with Replay Resistance. In *Proceedings of IEEE Symposium on Security and Privacy (S&P)*, 2023.

[96] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni. Generalizing From a Few Examples: A Survey on Few-shot Learning. *ACM Computing Surveys (CSUR)*, 53(3):1–34, 2020.

[97] J. Wu, Q. Liu, Z. Huang, Y. Ning, H. Wang, E. Chen, J. Yi, and B. Zhou. Hierarchical Personalized Federated Learning for User Modeling. In *Proceedings of ACM Web Conference (WWW)*, 2021.

[98] N. Wu, R. Cheng, S. Chen, and B. Han. Preserving Privacy in Mobile Spatial Computing. In *Proceedings of the Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV).*, 2022.

[99] Y. Wu, C. Shi, T. Zhang, P. Walker, J. Liu, N. Saxena, and Y. Chen. Privacy Leakage via Unrestricted Motion-Position Sensors in the Age of Virtual Reality: A Study of Snooping Typed Input on Virtual Keyboards. In *Proceedings of IEEE Symposium on Security and Privacy (S&P)*, 2023.

[100] Q. Yang, Y. Liu, T. Chen, and Y. Tong. Federated Machine Learning: Concept and Applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

[101] J. Yi, S. Choi, and Y. Lee. EagleEye: Wearable Camera-based Person Identification in Crowded Urban Spaces. In *Proceedings of ACM MobiCom*, 2020.

[102] X. Yin, Y. Zhu, and J. Hu. A Comprehensive Survey of Privacy-preserving Federated Learning: A Taxonomy, Review, and Future directions. *ACM Computing Surveys (CSUR)*, 54(6):1–36, 2021.

[103] F. Yu, A. S. Rawat, A. Menon, and S. Kumar. Federated Learning with Only Positive Labels. In *Proceedings of ICML*, 2020.

[104] T. Zhang, Q. Ji, Z. Ye, M. M. R. R. Akanda, A. T. Mahdad, C. Shi, Y. Wang, N. Saxena, and Y. Chen. SAFARI: Speech-Associated Facial Authentication for AR/VR Settings via Robust Vibration Signatures. In *Proceedings of ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2024.

[105] T. Zhang, Z. Ye, A. T. Mahdad, M. M. R. R. Akanda, C. Shi, Y. Wang, N. Saxena, and Y. Chen. FaceReader: Unobtrusively Mining Vital Signs and Vital Sign Embedded Sensitive Info via AR/VR Motion Sensors. In *Proceedings of ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2023.

[106] Y. Zhang, W. Hu, W. Xu, C. T. Chou, and J. Hu. Continuous Authentication Using Eye Movement Response of Implicit Visual Stimuli. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, 1(4):1–22, 2018.

[107] H. Zhu, W. Jin, M. Xiao, S. Murali, and M. Li. Blinkey: A Two-Factor User Authentication Method for Virtual Reality Devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(4):1–29, 2020.

[108] H. Zhu, M. Xiao, and M. Li. SoundLock: A Novel User Authentication Scheme for VR Devices Using Auditory-Pupillary Response. In *Proceedings of Network and Distributed System Security Symposium (NDSS)*, 2023.

[109] H. Zhu, J. Xu, S. Liu, and Y. Jin. Federated Learning on non-IID data: A Survey. *Neurocomputing*, 465:371–390, 2021.