When (Inter)actions Speak Louder Than (Pass)words: Task-Based Evaluation of Implicit Authentication in Virtual Reality

Woojin Jeon
Department of Electrical and Computer Engineering
Sungkyunkwan University
Suwon, Republic of Korea
dnwls0116@skku.edu

Chaejin Lim
Department of Electrical and Computer Engineering
Sungkyunkwan University
Suwon, Republic of Korea
chaejin98@skku.edu

Hyoungshick Kim
Department of Electrical and Computer Engineering
Sungkyunkwan University
Suwon, Republic of Korea
hyoung@skku.edu

***Abstract*—We present a practical implicit authentication system for Virtual Reality (VR) that uses natural interaction tasks—grabbing, pointing, and typing—as behavioral biometrics. The system extracts 221 features from head-mounted and controller sensors and is trained as a lightweight SVM-based binary classifier using data from legitimate users and a small set of reference users to simulate attacker behavior. In a 24-participant study, our system achieved strong authentication performance, with median Equal Error Rates (EERs) of 0.4% for grabbing, 2.6% for pointing, and 0.3% for typing. Designed for on-device deployment, it requires no GPU support, completes inference within 1 second, and maintains a compact model size under 0.2 MB, enabling efficient, real-time authentication on standalone VR headsets. Security evaluations with attacker-in-the-loop experiments across no-knowledge, shoulder-surfing, and video-replay conditions revealed clear trade-offs. Typing and pointing offered strong resistance to impersonation, while grabbing, despite high usability, was more vulnerable under video replay with a 23.8% attack success rate. These results demonstrate that secure, accurate, and real-time implicit authentication is feasible in VR, with task-specific characteristics enabling flexible deployment based on security and usability needs.**

***Index Terms*—User Authentication, Virtual Reality, Behavioral Biometrics**

I. Introduction

As Virtual Reality (VR) becomes more integrated into daily life through applications in education, healthcare, social interaction, and gaming [1]–[3], securing these personal devices is becoming increasingly important. This trend highlights the need for secure authentication and effective device locking mechanisms [4].

Despite the need for secure access, traditional authentication methods such as passwords, PINs, and biometrics are poorly suited for VR [5]. Even malicious applications with access only to built-in sensors on VR devices can potentially infer typed passwords, posing a serious threat to user credentials [6]. In addition, they break immersion, require user effort, and often force transitions between virtual and physical interfaces [7]. For instance, entering credentials without a physical keyboard is error-prone and awkward [8]. Biometric scans, though convenient, face limitations in VR. They often require additional sensors, which increases device cost and complexity [9], and users may resist them due to concerns about storing sensitive data on headsets without secure hardware [10]. These issues underscore the need for reliable, privacy-preserving alternatives. In VR, authentication must be seamless, low-friction, and secure even in adversarial conditions.

Prior work has explored various VR authentication methods. Explicit behavioral authentication systems require users to perform predefined gestures or movements [11]–[13], while others leverage physiological signals such as eye gaze or heart rate [14]–[16]. Although these approaches offer security benefits, they often compromise user experience or require specialized hardware. Moreover, most existing methods unrealistically assume access to behavioral data from unauthorized users (*i.e.*, potential attackers) during training—a condition rarely met in real-world deployments. This leads to impractical binary classification approaches that rely on predicting the behavior of unknown adversaries. In practice, model training should rely exclusively on legitimate user data. Furthermore, these methods inadequately address targeted attacks against willing victims, focusing primarily on user identification rather than comprehensive threat scenarios.

To address these limitations, we propose a practical implicit authentication framework that leverages natural user interactions (*e.g.*, grabbing, pointing, and typing) as behavioral biometrics. These interactions are already embedded in everyday VR use and require no additional effort, enabling passive and continuous authentication. Our system initially extracts 224 behavioral features from the Head-Mounted Display (HMD) and controller sensor streams, including motion trajectories, spatial relationships, and timing patterns. After conducting a feature importance analysis, we refined the dataset to 221 critical features and trained a lightweight Support Vector Machine (SVM) classifier for user verification. To address the absence of unauthorized user data during training, we utilize behavioral data from other users to represent potential imposters. Additionally, to mitigate the class imbalance between limited legitimate user samples, we apply SMOTE-based data augmentation [17].

We evaluate our system through a user study with 24 participants who performed three natural VR interaction tasks—grabbing, pointing, and typing—across temporally separated sessions spaced 30 minutes and 3 days apart. This setup enables analysis of behavioral consistency over time. Our lightweight SVM-based model achieves high authentication performance, with median Equal Error Rates (EERs) of 0.4% for grabbing, 2.6% for pointing, and 0.3% for typing under realistic training conditions.

The system is designed with practical deployment in mind. It completes inference in under one second for all tasks, finishes training in under 45 seconds, and requires no GPU support. With a compact model size of 0.2 MB and low memory usage, it is well-suited for on-device execution in standalone VR headsets, enabling real-time authentication without relying on cloud infrastructure.

This work is guided by the following research questions:

- **RQ1:** Can behavioral data from other users be effectively used to construct a two-class authentication model that outperforms conventional one-class approaches?
- **RQ2**: Can the three natural interaction tasks in VR environments (*e.g.*, grabbing, pointing, and typing) serve as effective implicit behavioral biometrics for user authentication, and which task is most effective?
- **RQ3**: How robust is the proposed authentication approach against observation-based attacks such as shoulder-surfing and recording in VR environments?

To address these questions, we make the following key contributions:

- We present a practical implicit authentication framework for VR that leverages natural user interactions—grabbing, pointing, and typing—as behavioral biometrics. Our system extracts 221 features from head-mounted and controller sensor data and employs a lightweight SVM classifier to achieve robust performance, with median EERs of 0.4% for grabbing, 2.6% for pointing, and 0.3% for typing. Designed for efficiency, the framework operates fully on-device without GPU support and completes inference in under one second, ensuring secure and unobtrusive VR authentication in real-world deployments.
- We reformulate the authentication task as binary classification by using behavioral data from a small group of reference users to represent unauthorized behavior during training. Although these users are excluded from evaluation, their inclusion enables the model to outperform one-class baselines by learning more discriminative decision boundaries without access to real attacker data.
- We evaluate the system's robustness against three realistic impersonation attacks—no-knowledge, shoulder-surfing, and video-replay—through attacker-in-the-loop experiments, providing the first empirical evaluation with real attackers in VR authentication.
- We release our anonymized dataset and full implementation at https://github.com/Jason-WJ96/implicit-vr-auth to promote transparency and reproducibility.

II. Background and Related Work

The difference between VR devices and traditional digital devices creates a demand for more suitable user authentication systems. Knowledge-based authentication methods in VR, such as PINs, patterns, or passwords, follow the theoretical foundations of traditional authentication methods. However, these tend to fall short in usability due to the different forms of user input by hand-held controllers. Riyadh *et al.* [8] evaluated user experience during VR authentication by comparing PIN and gesture, where PIN resulted in lower usability scores and required significantly longer authentication time. Additionally, the security of these traditional methods is reported to be more vulnerable when implemented in VR systems, enabling shoulder-surfing attacks due to their monotony [18]. As a result, researchers are now focusing on biometric-based authentication systems to deliver a more secure and seamless user experience.

Physiological biometric authentication methods explore various modalities to provide robust authentication schemes that do not disrupt the user experience. Brainwave responses [14], [19], Electrical Muscle Stimulation (EMS) [16], or auditory-pupillary responses [15] are recorded and utilized for authentication with high performance. However, these require additional sensors, along with the VR device, which reduces their practicality for real-world deployment. Shao *et al.* [20] address this issue by leveraging eye-tracking components embedded in VR headsets to authenticate users across different content types. However, it has only been demonstrated on passive content types, such as video and text. Moreover, eye-tracking hardware is typically found only in high-end VR devices and is rarely available in consumer-grade headsets.

Behavioral biometric authentication methods overcome this expense by leveraging features inherent to the VR device itself, capturing user behavior through specific tasks. Wang *et al.* [13], inspired by the earlier "Slide to Unlock" on mobile devices, designed a nodding gesture for user authentication. Rupp *et al.* [21] utilized 10 different and simple gestures (*e.g.*, wave, thumbs-up) into a sequence, aiming to substitute virtual PINs. Relatively simpler to collect and use compared to physiological biometrics, behavioral biometrics still face challenges such as limited scalability due to similarities between users, or balancing between authentication performance and input efficiency. Liebers *et al.* [22] poses another challenge by analyzing the drop in performance due to the lack of behavioral consistency over long-term periods.

Miller *et al.* [23] proposed a ball-throwing task for VR authentication based on user behavior, collecting data across multiple VR devices. Building on this work, they enhanced authentication performance using Siamese Neural Networks [24]. Li *et al.* [11] utilized motion forecasting via a Transformer-based model [25], achieving promising EER results. However, these studies overlook practical deployment concerns by using behavioral data from users who are treated as unauthorized during training, while the same users are also used in evaluation. This design inflates performance and fails to meet real-world constraints, as it relies on the unrealistic assumption that attacker behavior is known in advance.

While behavioral biometrics have been widely explored in mobile and desktop settings, their application in immersive VR remains underexplored. Teather *et al.* [26] evaluated VR pointing and grabbing interactions under ISO9241-9 [27], laying early groundwork for movement-based user modeling. Pfeuffer *et al.* [28] expanded on this by incorporating virtual keyboard typing tasks using MacKenzie *et al.*'s [29] standardized sentence set, and investigated head and hand motion features for user authentication.

Jeon *et al.* [30] investigated the feasibility of implicit authentication in VR using natural interaction tasks and a Random Forest classifier, based on data from 16 participants and without attacker-in-the-loop studies. Our current work advances this research in three major ways: (1) we optimize a 224-dimensional feature set via ablation, selecting 221 features for improved accuracy; (2) we systematically compare five classifiers, including two one-class models, and adopt a lightweight SVM model for real-time, on-device inference; and (3) we conduct attacker-in-the-loop experiments across three observation-based impersonation scenarios to assess security under realistic threat models. Based on a study with 24 participants, our results provide a comprehensive evaluation of authentication performance, security, and usability trade-offs across tasks. These contributions offer a deeper understanding of the practicality and deployment readiness of implicit authentication in immersive VR environments.

III. Threat Model

The attacker's objective is to gain unauthorized access to a legitimate user's VR device in order to compromise the user's privacy. This may involve stealing sensitive personal information (*e.g.*, account credentials, biometric data, or recorded video), making unauthorized online purchases using stored credit card information, or installing malware—all of which can result in significant harm to the user. To achieve this goal, the attacker must wear the legitimate user's VR headset and effectively impersonate the user to bypass the authentication system. We assume the attacker has physical access to the VR device, encounters no physical limitations in wearing the headset, and is aware of the existence of the VR authentication mechanism. This capability reflects a realistic scenario in which a close acquaintance or family member gains access to the user's device while visiting the user's home—for instance, when the legitimate user steps away briefly, such as to answer a phone call or use the restroom. Based on the amount and type of information available to the attacker, we consider the following attack scenarios.

No-Knowledge Attack. In this scenario, the attacker attempts to pass the authentication process without having observed or learned anything about how the legitimate user performs authentication. The attacker knows only the identity of the legitimate user and general physical characteristics, such as height. They must rely entirely on guesswork or inference to mimic the user's authentication behavior. This represents the weakest attacker model in terms of prior knowledge and preparation.

Shoulder-Surfing Attack. In this scenario, the attacker observes the legitimate user performing and successfully completing the authentication process at close range. Although the attacker can visually monitor the user's movements, the number of observation opportunities is limited to only two or three instances. When attempting to bypass the authentication system, the attacker must rely entirely on short-term memory and personal interpretation of the observed behavior. This attack reflects a realistic threat in shared environments, where an attacker may gain brief but direct visual access to the authentication process.

Video-Replay Attack. In this scenario, the attacker secretly records the legitimate user performing the authentication task using a smartphone or video camera from various angles (*e.g.*, front, side, and back). Unlike in the shoulder-surfing attack, the attacker can replay and analyze the recorded footage multiple times before attempting the attack. A practical example of this threat is a guest covertly recording the user during an initial visit to their home and then attempting the attack during a subsequent visit. This scenario represents a more powerful and intentional adversary model, as the attacker can study the user's detailed movement patterns and behaviors prior to launching the attack.

IV. System Design

Figure 1 presents the overall architecture of the proposed behavioral authentication system for VR environments. The system is composed of three core modules: a feature extraction module that derives representative behavioral features from sensor data collected during user interaction, a classifier training module that constructs user-specific authentication models based on the extracted features, and a real-time authentication module that verifies user identity by continuously authenticating the user's incoming behavioral data.

A. Feature Extraction Module

During user interaction with VR content, while wearing an HMD, Left Controller (LC), and Right Controller (RC), multiple embedded sensors continuously collect data reflecting the user's physical behavior. Specifically, each device collects three-dimensional positional coordinates (x, y, z), rotational orientation $(pitch, yaw, roll)$, linear velocities corresponding to position changes, and angular velocities derived from rotational movements. Additionally, the controllers collect button pressure values, ranging from 0 to 1, which indicate the degree to which each controller button is pressed. These continuous streams of raw sensor data are passed to the feature extraction module, transforming them into behavioral features that capture distinctive user movement and interaction patterns.

The feature extraction module computes a total of 224 features, categorized into four key behavioral dimensions.

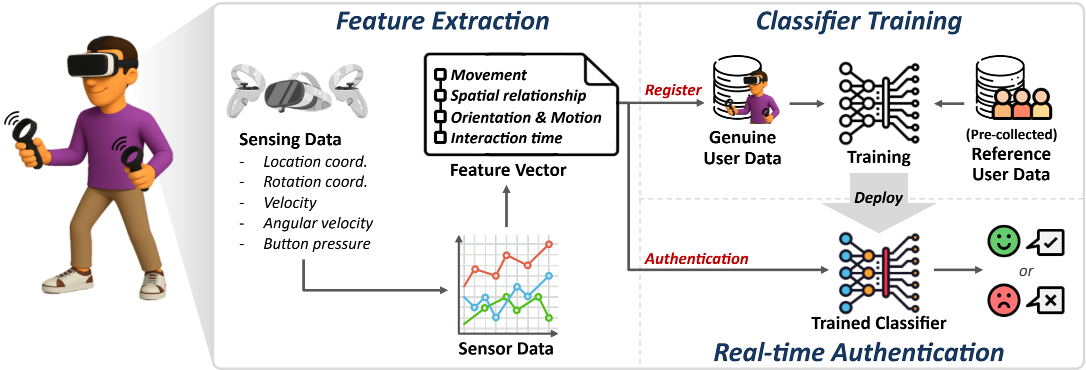


Fig. 1: Overview of the proposed behavioral authentication system for VR environments.

However, through ablation studies, we found that excluding the movement features slightly improved classification accuracy. Therefore, the final model used only 221 features, omitting the movement-related signals (see Section VI-B for details).

Movement Features. To quantify the extent of physical activity, the system calculates the total movement distance for each device (HMD, LC, and RC) by summing the Euclidean distances between consecutive 3D position samples over time. This approach captures the overall intensity of motion during a task, resulting in 3 features. While intuitively meaningful, these features were excluded in the final model, as they did not contribute to improved authentication performance in practice.

Spatial Relationship Features. For a detailed analysis of spatial positioning between devices, the system computes pairwise distances between devices (HMD–LC, HMD–RC, LC–RC), the vertical height of each device from the ground, and the relative horizontal distances from the controllers to the estimated body center (based on the HMD's position). For each of the eight spatial parameters, five statistical metrics (mean, median, minimum, maximum, and standard deviation) are extracted, yielding 40 features that capture spatial consistency, postural tendencies, and coordination patterns. These spatial measurements reveal how the user arranges and maneuvers their body and controllers within the virtual environment.

Device Orientation and Motion Features. To capture detailed motion patterns and characteristics of users, the system processes each device's rotational coordinates (pitch, yaw, roll), linear velocity, and angular velocity. The rotational coordinates are transformed into their sine and cosine representations to handle the periodic nature of rotation angles, ensuring continuity in the feature space. Then, five statistical measures are computed for each parameter, producing a total of 180 features that describe how the user rotates, maneuvers, and controls the devices during interaction.

Interaction Timing Feature. To represent the temporal aspect of user interaction, the system measures the total duration of button presses on the dominant-hand controller during interaction. This single feature captures how frequently and for how long the user engages with the input system, offering insight into their interaction style.

Feature extraction is performed locally on the VR device to minimize latency and ensure user data privacy. The resulting feature vectors are used in both the classifier training phase and the real-time authentication process.

B. Classifier Training Module

This module is responsible for constructing an authentication model that distinguishes between legitimate and unauthorized users based on behavioral features. Given that our authentication model is trained directly on the VR device, we deliberately focus on classical machine learning methods rather than deep learning approaches. On-device training imposes strict constraints on computational resources, memory usage, and training time. While powerful in many contexts, deep learning approaches typically require large-scale datasets, extensive training time, and specialized hardware accelerators, making them less suitable for real-time deployment on VR devices. In contrast, conventional machine learning models can be trained efficiently on limited data and within tight resource budgets while still achieving high levels of classification performance.

To enable robust classification between legitimate and unauthorized users, our authentication model is trained under a two-class setting. However, in real-world deployment, behavioral data from unauthorized users is typically unavailable at training time, as the VR device is initialized by a specific end user. To address this constraint, we adopt a proxy-based strategy using reference users—individuals whose behavioral data is collected offline in advance to simulate a range of unauthorized behaviors. These reference user profiles are preloaded onto the device before deployment.

To evaluate the effectiveness of this proxy-based training approach under realistic deployment constraints, we systematically evaluate and compare the performance of five machine learning classifiers. Specifically, we consider three binary classification models, Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR), which are trained using behavioral data from legitimate user and the preloaded reference users labeled as unauthorized. In addition, we include two one-class classification models, One-Class SVM (OCSVM) and One-Class K-Nearest Neighbor (OCKNN), which are trained solely on legitimate user data

TABLE I: Demographics of user study participants.

Gender			**Dominant hand**		
Female	13	(54.2%)	Right-handed	23	(95.8%)
Male	11	(45.8%)	Left-handed	1	(4.2%)
Age			**Heights**		
20 - 24	4	(16.7%)	150 - 159.9 cm	7	(29.2%)
25 - 29	16	(66.7%)	160 - 169.9 cm	6	(25%)
30 - 34	2	(8.3%)	170 - 179.9 cm	6	(25%)
35 - 39	2	(8.3%)	180 - 189.9 cm	4	(16.6%)
40 and more	0	(0%)	Over 190 cm	1	(4.2%)

and aim to detect anomalies without relying on unauthorized user samples. The hyperparameter configurations for each classifier are as follows. RF uses 100 estimators trained on GINI index. SVM uses the RBF kernel with the coefficient for regularization as 1.0. LR penalizes the L_2 norm with 1.0 as the coefficient for regularization with maximum iteration up to 100 times. OCSVM also utilizes the RBF kernel. OCKNN clusters samples with the number of neighbors as 5, selecting the neighbors based on the largest similarity scores calculated by average. This comparison enables us to assess whether our proxy-based two-class formulation offers measurable advantages over one-class modeling approaches commonly used in scenarios with limited negative data. Through this evaluation, we seek to determine not only which classifier performs best in terms of authentication accuracy, but also whether the use of reference users meaningfully improves model robustness and discriminative capability in practical VR environments.

While users provide only limited behavioral data during initial enrollment, the inclusion of pre-collected reference user data enables the model to learn discriminative boundaries despite the class imbalance. All training is performed locally on the VR device, ensuring that no raw behavioral data is transmitted externally. This design preserves user privacy while supporting secure and personalized authentication.

C. Real-time Authentication Module

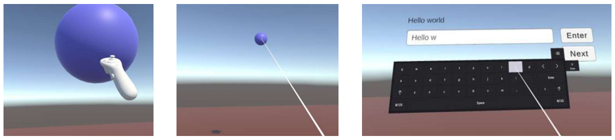
The real-time authentication module operates continuously during VR usage to validate user identity in coordination with the feature extraction module and trained authentication model. As sensor data streams from the HMD and controllers, it is processed into feature vectors, which the trained model then evaluates to produce authentication scores. By comparing each score against a configurable threshold, the authentication module decides whether to accept or reject the current user.

V. User Study

A. Participants

A total of 24 participants were recruited from a university. As shown in Table I, the sample included 11 male and 13 female participants, resulting in a slightly female-skewed gender distribution. The majority of participants (66.7%) fell within the 25-29 age range, with smaller subsets in the 20-24 (16.7%), 30–34 (8.3%), and 35–39 (8.3%) age brackets, and none over 40 years of age (mean: 27 years, SD: 3.76). Participants exhibited a broad range of heights, with the largest proportion (29.2%) falling within 150–159.9 cm, followed by 25.0% each in the 160–169.9 cm and 170–179.9 cm ranges, 16.6% in 180–189.9 cm, and one participant exceeding 190 cm (mean: 168 cm, SD: 10.3). Regarding handedness, 95.8% of participants reported being right-handed, with only one individual (4.2%) identifying as left-handed.

Fig. 2: Study participant using a Meta Quest Pro headset.



(a) Grabbing (b) Pointing (c) Typing

Fig. 3: Three natural interaction tasks in VR.

B. Apparatus

As shown in Figure 2, the user study was conducted using the Meta Quest Pro, a virtual reality device composed of a head-mounted display and two motion-tracked controllers. To facilitate data collection and guide participants through the experiment, a custom-built VR application was developed using Unity 3D and programmed in C#. This application was executed on a Windows 10 PC featuring an Intel Core i9-12900F CPU, 32 GB of memory, and an NVIDIA GeForce RTX 3070 GPU. The VR headset was connected to the PC via a wired link, which ensured consistent application performance and minimal latency throughout the experimental sessions. During usage, the Meta Quest Pro's onboard sensors recorded interaction data at a sampling frequency of 100 Hz, which we selected as a balanced trade-off between temporal resolution and computational efficiency. While prior studies have used sampling rates ranging from 45 to 200 Hz [13], [23], 100 Hz provided sufficient granularity to capture meaningful behavioral patterns during natural VR interactions.

C. VR Interaction Tasks

To evaluate the effectiveness of the proposed authentication system, we employ the three core VR interaction tasks explored by Pfeuffer *et al.* [28]: grabbing, pointing, and typing (see Figure 3). These tasks represent essential modalities through which users naturally engage with VR environments, whether manipulating objects, selecting distant elements, or entering text.

Grabbing Task. As shown in Figure 3a, this task simulates direct physical manipulation of virtual objects. A virtual object is captured by moving the controller to the object boundary and pressing a specified button. Once engaged, the object becomes attached to the controller and follows its motion until the button is released, closely paralleling real-world grasping behaviors. In our experiment, spherical virtual objects were placed along the circumference of a circle with a diameter of 0.7 m, centered 1.5 m above the ground, and 0.5 m in front of the participant. Each object appeared one at a time in a fixed sequence, with 13 targets per session. Participants completed 15 sessions for each of two object sizes: 0.15 m and 0.3 m in diameter.

Pointing Task. This task simulated ray-based interactions with distant virtual objects. Participants pointed a ray emitted from the controller to align with a target sphere and pressed a button to select it. The spheres were arranged in a 2 m diameter circle at a height of 1.5 m, with 13 targets appearing sequentially in each session. Each participant completed 15 sessions per distance condition (2 m and 4 m), enabling us to examine performance across different interaction ranges.

Typing Task. As shown in Figure 3c, this task simulates text input via a virtual keyboard. The user aligns the controller's ray with a target key and presses the designated button to input the corresponding letter. This interaction supports critical functions such as communication, form filling, and command input in VR applications. In our experiment, participants entered text using a virtual keyboard positioned 3.75 m in front of them. A target sentence was displayed for each session, and participants selected keys by aiming with a ray pointer and pressing a button. Each participant completed 20 sessions, typing a different pre-defined sentence in each.

Through these three interaction tasks, we investigate whether naturally occurring user behaviors in typical VR scenarios can serve as reliable indicators for identity verification.

D. Procedure

Before conducting the study, all participants were provided with a detailed explanation of the study's objectives, procedures, and their rights as participants. After completing a consent form and demographic questionnaire, participants were guided through an orientation session, where they learned how to properly wear the VR headset and interact with our proposed authentication system. Participants conducted a total of three studies based on the designed study design. The first two studies were conducted on the same day, with a 30-minute break in between to introduce a short-term temporal separation between data recordings. During the break, participants removed the VR headset and controllers, and re-equipped them before starting the second study. The third study was conducted at least three days after the second one, in order to evaluate the consistency of behavioral patterns over a longer-term interval. In each study, participants completed the tasks in the following order: grabbing, pointing, and typing. This temporal separation allowed us to assess both short-term and long-term stability of behavioral biometric patterns.

E. Ethical Considerations

Our user study was approved by the Institutional Review Board (IRB) at our institution. All participants were recruited voluntarily and provided informed consent before participating in the study. The purpose of the study, procedures, and data handling practices were clearly explained to each participant, and they were informed of their right to withdraw at any time without penalty. The tasks involved standard VR interactions with no known physical or psychological risks. To further mitigate discomfort, participants were allowed to take breaks as needed and were debriefed after the study was completed.

VI. Evaluation

A. Evaluation Framework

Dataset Overview. We collected data from 24 participants, each of whom completed three separate sessions. In each session, 30 samples were collected per participant for the *grabbing* and *pointing* tasks, and 20 samples for the *typing* task. This resulted in 2,160 samples (= 30 samples × 3 studies × 24 participants) for grabbing and pointing, and 1,440 samples for typing.

Reference Users. To enable two-class training despite the absence of real adversaries, we selected four *reference users* based on their height ranks among the 24 participants: the 4th (≈15%), 10th (≈40%), 16th (≈65%), and 22nd (≈90%). These individuals were used exclusively for training as proxy imposters and were excluded from all evaluations.

Evaluation Scenarios. We designed two scenarios to assess the system's robustness under different temporal conditions. For both scenarios, authentication performance was independently evaluated for each interaction task—grabbing, pointing, and typing—to compare their relative discriminative power.

- Scenario 1: The model is trained using the authorized user's data from study 1 as positive samples and reference user data as negative samples. Evaluation is performed on the authorized user's data from study 2 and 3, along with study 2 and 3 data from the remaining 19 participants as unseen unauthorized users.
- Scenario 2: The model is trained using the authorized user's data from study 1 and 2, and the same reference user data as negative samples. Evaluation is performed on study 3 data from both the authorized user and the remaining participants.

Scenario 2 simulates a practical deployment case where the model can be updated over time with new data. It allows us to assess whether incorporating temporally adjacent training data (study 2) helps the model generalize to longer-term behavioral variations (study 3).

Evaluation Metrics. System performance was evaluated using three standard authentication metrics: Equal Error Rate (EER), Receiver Operating Characteristic (ROC) curve with Area Under the Curve (AUC), and Precision-Recall (PR) curve with Average Precision (AP). EER indicates the threshold at which false acceptance and false rejection rates are equal, offering

TABLE II: Median EER of different classification models in Scenario 1, where models are trained on study 1 data and tested on study 2 and 3 data.

Task	RF	SVM	LR	OCSVM	OCKNN
Grabbing	17.9 (17.7)	**3.9** (11.3)	7.5 (9.6)	6.7 (11.8)	6.0 (12.0)
Pointing	10.5 (16.0)	**7.4** (11.1)	10.0 (13.8)	11.1 (14.2)	11.1 (16.4)
Typing	7.3 (16.7)	**3.6** (5.6)	4.5 (8.8)	5.0 (6.5)	5.1 (6.6)

* *Values in parentheses indicate the interquartile range (IQR)*

TABLE III: Median EER of different classification models in Scenario 2, where models are trained on study 1 and 2 data and tested on study 3 data.

Task	RF	SVM	LR	OCSVM	OCKNN
Grabbing	6.8 (13.2)	**0.8** (1.8)	6.1 (6.3)	6.2 (7.3)	5.9 (8.0)
Pointing	6.6 (8.7)	**2.5** (7.0)	5.7 (7.7)	10.0 (16.5)	9.3 (16.5)
Typing	0.9 (7.7)	**0.8** (3.4)	2.4 (6.0)	6.1 (7.3)	6.6 (7.4)

* *Values in parentheses indicate the IQR.*

a single, interpretable measure of system accuracy. ROC-AUC captures the model's overall discriminative ability across thresholds, while PR-AP emphasizes performance under class imbalance, making it particularly suitable given the limited size of the legitimate user class. Together, these metrics provide a robust and comprehensive evaluation framework for authentication tasks.

B. Performance Evaluation

Comparison Across Classification Models. Tables II and III compare the median EERs of five classification models under two training scenarios. Results are presented for both Scenario 1 (single-session enrollment) and Scenario 2 (multi-session enrollment). We report median rather than mean to reduce sensitivity to outliers and better reflect typical authentication performance. In both scenarios, SVM consistently achieved the lowest EERs across all tasks.

In Scenario 1, SVM achieved EERs of 3.9% (grabbing), 7.4% (pointing), and 3.6% (typing), outperforming all other classifiers. With additional user data in Scenario 2, performance improved further: 0.8% (grabbing), 2.5% (pointing), and 0.8% (typing). These results highlight the effectiveness of SVM and the benefit of multi-session training. The task-level EERs of SVM ranged from 3.6%–7.4% in Scenario 1, and improved to 0.8%–2.5% in Scenario 2. In contrast, RF and LR exhibited relatively high error rates in Scenario 1. RF achieved EERs from 7.3% (typing) to 17.9% (grabbing), while LR ranged from 4.5% (typing) to 10.0% (pointing). However, both models demonstrated notable improvements in Scenario 2. RF reduced its EERs to a range of 0.9% (typing) to 6.8% (grabbing), and LR achieved EERs from 2.4% (typing) to 6.1% (grabbing), indicating better generalization with increased behavioral data.

TABLE IV: Comparison of training time, inference time, and model size across models in Scenario 2

Performance	Task	RF	SVM	LR
Training time (sec.)	**Grabbing**	12.3	**11.5**	**11.5**
	Pointing	14.0	**13.2**	13.8
	Typing	22.7	**21.2**	22.2
Inference time (sec.)	**Grabbing**	**0.2**	**0.2**	**0.2**
	Pointing	**0.2**	**0.2**	**0.2**
	Typing	**0.5**	**0.5**	**0.5**
Model size (MB)	**Grabbing**	0.1	0.2	**0.003**
	Pointing	0.1	0.2	**0.003**
	Typing	0.1	0.2	**0.003**

In contrast, the one-class models consistently underperformed relative to binary classifiers. In Scenario 1, OCKNN yielded EERs from 6.0% (grabbing) to 16.4% (pointing), while OCSVM ranged from 5.0% (typing) to 11.1% (pointing). Even with additional training data in Scenario 2, both models remained less competitive. OCKNN recorded EERs between 6.6% (typing) and 9.3% (pointing), and OCSVM ranged from 6.1% (typing) to 10.6% (pointing), highlighting their limited effectiveness despite multi-session input.

These results reinforce the strength of the SVM-based binary classification framework, with reference users, in achieving low EERs across various tasks and scenarios. However, to validate its practical suitability for real-world deployment in VR environments, we further evaluated the binary classifiers (SVM, RF, and LR) in terms of resource efficiency, including training time, inference latency, and model size (see Table IV). For the grabbing task, SVM and LR required 11.5 seconds to train, slightly faster than RF at 12.3 seconds. In the pointing task, SVM was the fastest, completing the task in 13.2 seconds, followed by LR (13.8 seconds) and RF (14.0 seconds). A similar trend was observed in the typing task, with SVM (21.2 seconds) again training faster than LR (22.2 seconds) and RF (22.7 seconds).

Inference latency was identical across all models, at 0.2 seconds for both grabbing and pointing tasks, and 0.5 seconds for typing. These consistent sub-second response times demonstrate that computational latency is negligible for real-time authentication.

Model size varied between classifiers, with LR being exceptionally compact at 0.003 MB for all tasks. SVM models were substantially larger (0.2 MB), while RF maintained moderate sizes (0.1 MB). Despite its larger size, SVM maintained the same inference speed as the other models while delivering superior authentication accuracy. Based on these combined performance metrics, we selected SVM as the primary model for subsequent evaluations.

Answer to RQ1: Our two-class SVM model, trained with behavioral data from other users, significantly outperformed one-class baselines across all tasks and scenarios, achieving median EERs as low as 0.8%. While one-class models showed limited improvement even with multi-session training, our approach generalized well, validating the effectiveness of using proxy users for two-class authentication.

TABLE V: Leave-one-feature-category-out ablation. “All” = all features; “M/S/O/I” = Movement, Spatial, Orientation, Interaction.

Scenario	Task	All	w/o M	w/o S	w/o O	w/o I
Scenario 1	Grabbing	**3.9** (11.3)	**3.9** (11.1)	5.1 (14.1)	10.4 (14.3)	**3.9** (11.2)
	Pointing	7.4 (11.1)	7.6 (9.1)	7.9 (14.9)	6.8 (8.0)	**6.5** (8.9)
	Typing	3.6 (5.6)	3.6 (5.3)	**3.4** (8.2)	5.1 (5.6)	3.9 (5.0)
Scenario 2	Grabbing	0.8 (1.8)	**0.6** (2.0)	1.8 (3.4)	6.3 (11.0)	0.8 (1.7)
	Pointing	**2.5** (7.0)	2.6 (7.4)	3.2 (6.5)	3.2 (10.1)	**2.5** (8.2)
	Typing	0.8 (3.4)	**0.7** (3.4)	3.0 (7.7)	**0.7** (3.5)	**0.7** (5.2)

* *Values in parentheses indicate the IQR.*

Feature Importance Analysis. To assess the contribution of each feature category to authentication performance, we conducted a *leave-one-feature-category-out* ablation study (see Table V) and applied SHAP [31] analysis to the trained SVM models. Figure 4 presents the SHAP value distributions of the four feature categories (Movement, Spatial, Orientation, Interaction) across the three tasks (Grabbing, Pointing, Typing). Higher SHAP values indicate that a feature increases the likelihood of the model predicting the positive class (authorized user), whereas lower values decrease it, favoring the negative class (unauthorized user). Due to class imbalance, with most samples belonging to the negative class, the baseline SHAP values are generally close to zero. Therefore, a higher SHAP value for a given feature category suggests greater influence in distinguishing between the two classes and indicates its importance in the model's decision-making process.

Overall, both methods consistently identified orientation features as the most impactful. These features exhibited the highest SHAP values across all tasks, with their influence most pronounced in the Grabbing task—removal of orientation features led to a significant increase in EER of 10.4% in Scenario 1 and 6.3% in Scenario 2. These features—including angular velocity and rotational components (pitch, yaw, roll)—captured fine-grained, user-specific motion patterns.

In contrast, the remaining feature categories reflected task-specific trends. In the Pointing task, models relied more on spatial features than on movement features, as the task requires minimal physical motion and primarily involves ray casting. This is evidenced by higher EERs when spatial features were removed (7.9% in Scenario 1, 3.2% in Scenario 2) and by higher SHAP values compared to movement features.

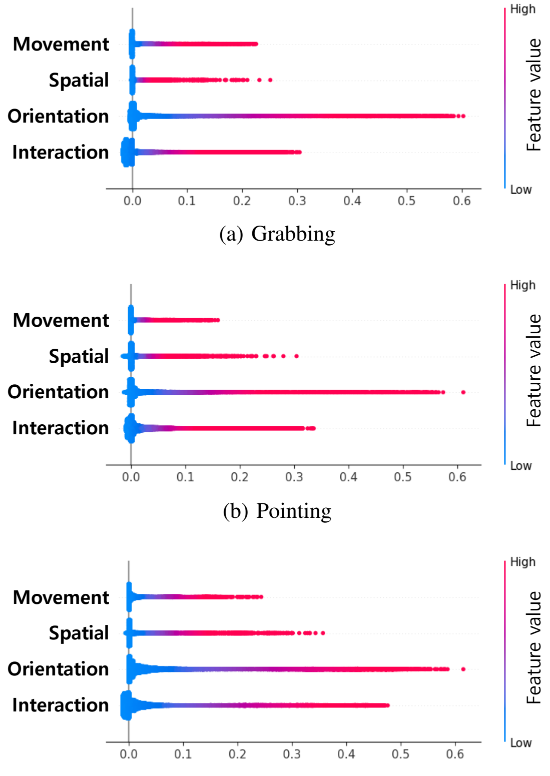
In the Grabbing and Typing tasks, distinguishable SHAP values indicate a substantial reliance on interaction features. However, their removal did not affect EER, suggesting that their influence may be redundant with other features or captured through non-linear interactions with other cues.

Movement features were consistently the least important, with low SHAP values and negligible impact on EERs across all tasks. Based on these findings, we excluded the three movement features from the final model, resulting in a 221-dimensional feature vector used in all subsequent experiments.

Comparison Across Three Tasks. To evaluate the relative effectiveness of different interaction modalities in VR authentication, we analyzed the performance of grabbing, pointing, and typing across two training scenarios using our final 221-feature configuration (excluding movement features, as determined by the ablation results in Table V). Task-level performance varied depending on the temporal scope of training. In Scenario 1, typing achieved the lowest median EER (3.6%), followed by grabbing (3.9%) and pointing (7.6%). In Scenario 2, grabbing achieved the lowest EER (0.6%), followed by typing (0.7%) and pointing (2.6%).

Figures 5 and 6 compare task separability using AUC and average precision (AP). In Scenario 1, typing showed the highest AUC (0.99) and AP (0.89), while grabbing (AUC: 0.95, AP: 0.79) and pointing (AUC: 0.95, AP: 0.76) yielded lower precision scores. In Scenario 2, grabbing achieved near-perfect performance (AUC: 1.00, AP: 0.96), aligning with its low EER and narrow IQR. Typing remained strong (AUC: 0.99, AP: 0.92), and pointing showed modest improvement (AUC: 0.95, AP: 0.82).

We then examined user-level variability using the IQR of EERs, as shown in Table V. In Scenario 1, grabbing exhibited the largest IQR (11.1%), indicating greater variation in user behavior. Typing showed the smallest IQR (5.3%), suggesting more consistent patterns across users. In Scenario 2, variability decreased across all tasks, with grabbing demonstrating the most substantial improvement—its IQR dropped to 2.0%, the

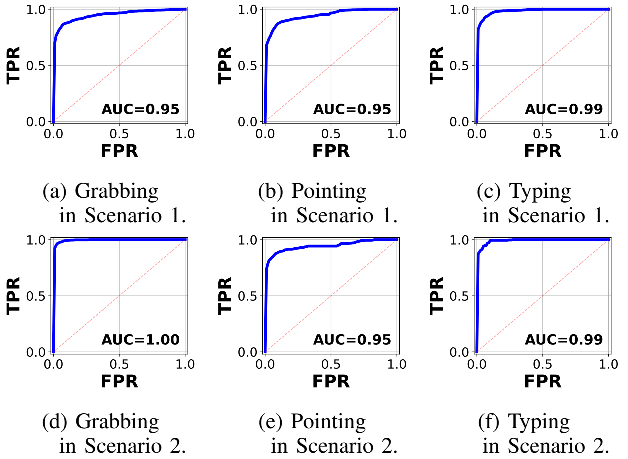


(a) Grabbing

(b) Pointing

(c) Typing

Fig. 4: SHAP value distributions for four feature categories.



(a) Grabbing in Scenario 1.
(b) Pointing in Scenario 1.
(c) Typing in Scenario 1.
(d) Grabbing in Scenario 2.
(e) Pointing in Scenario 2.
(f) Typing in Scenario 2.

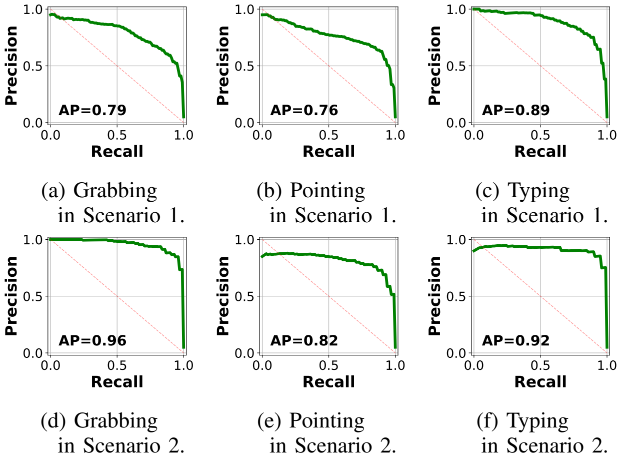
Fig. 5: Comparison of ROC curve and AUC across three tasks.

lowest among the tasks. Typing and pointing followed with IQRs of 3.4% and 7.4%, respectively.

These differences reflect the structural characteristics of each task. Typing involved constrained, repetitive, ray-based key selections, likely contributing to its stability even with limited training. Grabbing, in contrast, required full-hand motions and variable object placements, leading to higher variability in Scenario 1 but becoming more consistent with additional training in Scenario 2. Pointing, though mechanically simpler, involved targeting spheres at varying depths, which may have introduced motor inconsistencies across users.

Answer to RQ2: All three VR interaction tasks supported implicit authentication, but their effectiveness varied by task structure and training scope. In Scenario 1 (single-session), typing showed the best performance, with the lowest EER (3.6%), the highest AUC (0.99), and the smallest IQR (5.3%). In Scenario 2 (multi-session), grabbing became the most reliable task, achieving the lowest EER (0.6%) and a perfect AUC (1.00), followed by typing (EER: 0.7%, AUC: 0.99). Pointing remained less stable across users, with a higher EER (2.6%) and IQR (7.4%) compared to other tasks, indicating that it may require user-specific calibration or adaptation to achieve consistent performance.

Effects of Data Augmentation. To assess whether data augmentation can mitigate the class imbalance caused by limited legitimate user samples, we applied two techniques—SMOTE [17] and VAE [32]—to generate synthetic data for the authorized user class. The detailed configurations for each technique is as follows. SMOTE generates synthetic samples using 5 nearest neighbors with the training data shuffled prior to synthesis. VAE is implemented as a symmetric encoder-decoder structure with layer dimensions of 224-128-64-16-64-128-224, using ReLU as the activation function. It is trained for 1000 epochs with the Adam optimizer and a learning rate of 0.001. Their effectiveness was compared against a model trained without any augmentation. Table VI summarizes the results across three training scenarios.



(a) Grabbing in Scenario 1.
(b) Pointing in Scenario 1.
(c) Typing in Scenario 1.
(d) Grabbing in Scenario 2.
(e) Pointing in Scenario 2.
(f) Typing in Scenario 2.

Fig. 6: Comparison of PR curve and AP across three tasks.

TABLE VI: Median EERs across augmentation strategies; "No Aug." refers to training without augmentation.

Scenario	Task	No Aug.	SMOTE	VAE
Scenario 1	Grabbing	**3.9** (11.1)	4.8 (8.8)	8.0 (16.5)
	Pointing	7.6 (9.1)	**5.6** (6.8)	10.2 (14.3)
	Typing	**3.6** (5.3)	5.3 (3.8)	5.5 (5.3)
Scenario 2	Grabbing	0.6 (2.0)	**0.4** (3.2)	1.3 (3.3)
	Pointing	**2.6** (7.4)	**2.6** (5.5)	4.3 (10.0)
	Typing	0.7 (3.4)	**0.3** (5.3)	1.7 (5.3)

* *Values in parentheses indicate the IQR.*

In Scenario 1, augmentation showed mixed results. While SMOTE improved performance in pointing (EER reduced from 7.6% to 5.6%), it led to higher EERs in grabbing and typing. VAE was generally less effective, resulting in increased EERs across all tasks. These results suggest that synthetic data generated from sparse enrollment samples may not adequately capture intra-user variability, and in some cases, may introduce artifacts that hinder model generalization. Given these trade-offs, we recommend avoiding augmentation for short-term or single-session training settings.

In contrast, Scenario 2 revealed clear benefits of data augmentation. SMOTE consistently improved performance, reducing the EER from 0.6% to 0.4% for grabbing, and from 0.7% to 0.3% for typing, without degrading accuracy in pointing. VAE, however, remained less effective, increasing EERs across all tasks. These results highlight that SMOTE is a practical data augmentation strategy when sufficient behavioral diversity is present in the training data.

Ablation Study on Reference User Composition. To assess how the composition of reference users affects authentication performance, we conducted an ablation study under Scenario 2 using SMOTE. Starting from the *Base Group* of four reference users (see Section VI-A), we constructed four subgroups (*Subgroup1* to *Subgroup4*) by removing one user at a time

TABLE VII: Median EER across reference user groups using SMOTE in Scenario 2.

Task	Base Group	Subgroup1	Subgroup2	Subgroup3	Subgroup4
Grabbing	**0.4** (3.2)	1.2 (3.7)	1.1 (3.7)	0.6 (3.1)	0.5 (2.1)
Pointing	2.6 (5.5)	3.0 (8.4)	**2.3** (6.3)	3.2 (5.4)	3.4 (5.5)
Typing	**0.3** (5.3)	1.3 (5.2)	0.5 (3.0)	**0.3** (2.2)	0.5 (5.8)

* *Values in parentheses indicate the IQR.*

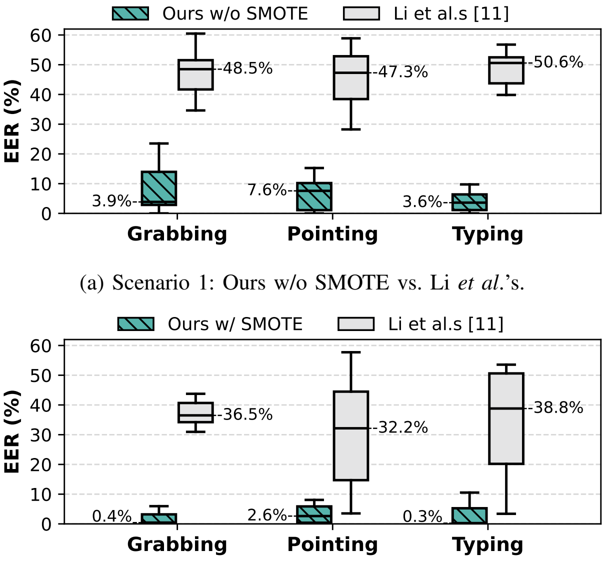
in ascending height order. Each subgroup consisted of three users, with all other settings remaining the same.

As shown in Table VII, performance varied across subgroups and tasks. For *grabbing*, the Base Group yielded the lowest median EER (0.4%), but Subgroup3 (0.6%) and Subgroup4 (0.5%) also maintained low error rates. Interestingly, for *pointing*, Subgroup2 slightly outperformed the Base Group (2.3% vs. 2.6%), while the other subgroups exhibited higher EERs. For *typing*, both the Base Group and Subgroup3 achieved the best results (0.3%), with larger degradation observed in Subgroup1 (1.3%).

These results indicate that the Base Group of four reference users generally yielded the best authentication performance across tasks. Although Subgroup2 showed a slight improvement in pointing (2.3% vs. 2.6%), most other subgroups underperformed, particularly in grabbing and typing. This suggests that reference user composition can significantly impact generalization, and that all tasks—grabbing, pointing, and typing—are sensitive to changes in the behavioral diversity of the reference set. Selecting users with consistent or representative interaction patterns is therefore critical for robust model performance.

Comparison with State-of-the-Art. We compared our method with the Transformer-based approach proposed by Li *et al.* [11], which utilizes only the temporal positions of the user's dominant hand for authentication. Evaluations were performed on our full dataset across grabbing, pointing, and typing tasks, measuring both authentication performance and resource efficiency.

As shown in Figure 7, our method significantly outperformed Li *et al.*'s across both training scenarios. In Scenario 2, EERs further decreased from 36.5% to 0.4% (grabbing), 32.2% to 2.6% (pointing), and 38.8% to 0.3% (typing). Although Li *et al.* reported EERs below 10% in their original paper [11], their method performed substantially worse on our dataset. This gap can be attributed to fundamental differences in task design, data diversity, and modeling assumptions. Li *et al.*'s method is tailored to a structured ball-throwing task and relies heavily on temporally aligned hand trajectories and future motion prediction using a Transformer-based architecture. Such an approach assumes highly regular behavior patterns, which may hold in constrained environments but fail to generalize to more varied settings. In contrast, our dataset includes three distinct natural VR interaction tasks—grabbing, pointing, and typing—each with unique motor demands and behavioral variability. By leveraging a richer set of multi-sensor features (from head and controllers), our method captures individualized behavior patterns more effectively. This broader representational capacity enables stronger generalization across ecologically valid and behaviorally diverse scenarios.



(a) Scenario 1: Ours w/o SMOTE vs. Li *et al.*'s.

(b) Scenario 2: Ours w/ SMOTE vs. Li *et al.*'s.

Fig. 7: EER comparison across tasks between our method and Li *et al.*'s method [11].

TABLE VIII: Comparison of training time, inference time, and memory usage between our method and Li *et al.*'s method [11].

Method	Task	Time (sec.)		Memory Usage (MB) (CPU ∣ GPU)	
		Training	Inference	Training	Inference
Li *et al.*'s [11]	Grabbing	9,684.6	481.4	4.9 ∣ 6,717	50.8 ∣ 6,719
	Pointing	10,229.3	491.2	4.9 ∣ 6,761	100.8 ∣ 6,773
	Typing	30,261.0	1,525.2	14.3 ∣ 6,704	326.9 ∣ 6,704
Ours (w/ SMOTE)	Grabbing	21.5	0.4	14.5 ∣ 0	1.3 ∣ 0
	Pointing	23.4	0.4	14.7 ∣ 0	1.6 ∣ 0
	Typing	43.4	0.9	15.2 ∣ 0	3.5 ∣ 0

In addition to accuracy, our method is significantly more efficient. As shown in Table VIII, training time was reduced from over 9,000 seconds to under 45 seconds across all tasks, and inference time from over 480 seconds to under one second. Unlike Li *et al.*'s approach, which required over 6 GB of GPU memory for both training and inference, our method operates entirely on the CPU with a small memory footprint. The model size also remains under 0.2 MB, making it suitable for deployment on resource-constrained VR devices.

These findings demonstrate the practicality of our method for real-time, on-device deployment in VR environments, providing both high authentication accuracy and exceptional efficiency.

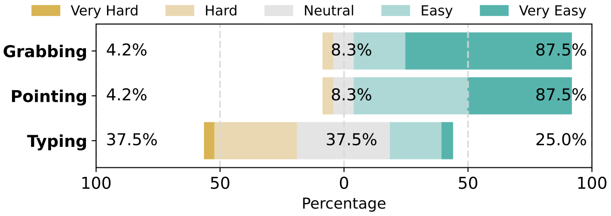


Fig. 8: Distribution of ease of use scores across three tasks.

C. Usability Evaluation

To assess the usability of the three VR interaction tasks, participants were asked to rate the ease of performing each task on a 5-point Likert scale, ranging from 1 (*Very Hard*) to 5 (*Very Easy*). The distribution of responses is presented in Figure 8.

The results demonstrate a clear disparity in perceived usability across tasks. Both grabbing and pointing were rated as highly usable, with 87.5% of participants rating these tasks as either "Easy" or "Very Easy." Notably, 66.7% of participants specifically classified grabbing as "Very Easy." In contrast, typing received substantially lower usability ratings, with only 25% of participants rating it as "Easy" or "Very Easy," while 75% found it difficult (37.5% rated it as "Very Hard" and 37.5% as "Hard"). To determine the statistical significance of these perceived differences, we conducted Mann–Whitney U tests with Bonferroni correction for multiple comparisons. No significant difference was observed between grabbing and pointing ($p = 0.4517$), indicating comparable usability. However, both were rated significantly easier than typing, with statistically significant differences observed in the comparisons between grabbing and typing ($p < 0.0001$) and between pointing and typing ($p < 0.0001$).

These findings demonstrate that VR interactions involving gross motor movements—such as grabbing and pointing—are perceived as more intuitive and less cognitively demanding by users. In contrast, typing, which requires precise finger articulation and mid-air character selection, poses substantially higher cognitive and motor challenges in virtual environments. This usability differential has important implications for the design of authentication systems that balance security with user experience.

D. Security Evaluation

Before presenting our security evaluation, we provide a summary of the threat model described in Section III. We consider three realistic attack scenarios, each varying in the attacker's level of knowledge and preparation. In the no-knowledge attack, the attacker has no prior exposure to the legitimate user's authentication behavior and relies solely on guesswork. In the shoulder-surfing attack, the attacker visually observes the user performing authentication a few times and attempts to mimic the observed behavior from memory. Lastly, in the video-replay attack, the attacker records the authentication process and analyzes the footage multiple times before launching the attack. These scenarios reflect a range of adversarial capabilities to evaluate the robustness of our proposed authentication system under various threat conditions.

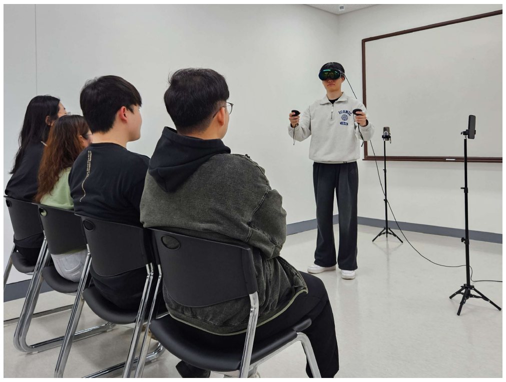


Fig. 9: *Shoulder-surfing* attack setup, where the attacker observes the victim's task performance at close range while multi-angle cameras record the session.

Experimental Setup. To assess the system's resistance to targeted attacks, we performed a dedicated experiment involving six participants from the original user pool described in Section V. Two participants (1 male, 1 female) were designated as victims, selected from height ranges of 170–179.9 cm and 150–159.9 cm, respectively. Four others, including two males and two females, served as attackers. For each victim, a male–female pair was chosen with a similar height (within 5 cm), and another pair with a notably different height (at least 10 cm apart). Attackers shorter than their assigned victims were provided a 6- to 10-cm footstool to reduce height discrepancies.

Each attacker performed three impersonation attack types in sequence. In the *no-knowledge* attack, they were given only the victim's height and completed 10 trials each of the grabbing, pointing, and typing tasks. In the *shoulder-surfing* attack, attackers observed the victim perform each task four times at close range, while multi-angle video recordings (front, side, and rear) captured the session (see Figure 9). They then conducted 10 impersonation trials per task. In the *video-replay* attack, attackers reviewed the recordings for 30–60 minutes before performing 10 informed impersonation trials per task.

Attack effectiveness was measured using the attack success rate, defined as the percentage of trials in which an attacker was mistakenly authenticated as the victim. This metric reflects the system's resilience against adversaries with increasing levels of behavioral knowledge.

Figure 10 summarizes attack success rates across three adversarial scenarios. While pointing and typing are highly resistant to impersonation, the grabbing task is susceptible to observation-based attacks.

No-Knowledge Attack. Out of 240 total attempts, four resulted in false acceptances (1.7%). Most occurred in the grabbing task (3.8%), followed by a single success in pointing

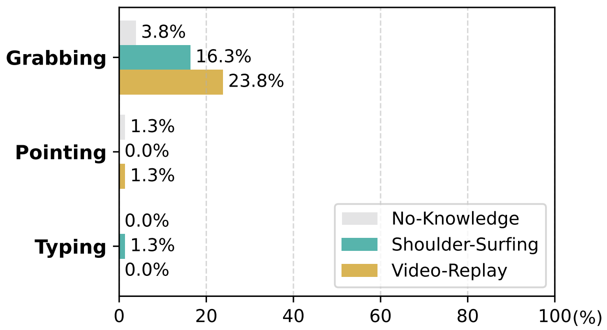


Fig. 10: Attack success rates by task under three adversarial scenarios: no-knowledge, shoulder-surfing, and video-replay.

(1.3%). Typing was unaffected (0%). Despite the overall low success rate, the results indicate that grabbing remains vulnerable even to uninformed attackers. This may be due to limited inter-user variability in grabbing motions, which makes it harder for the model to learn distinctive, user-specific boundaries.

Shoulder-Surfing Attack. The attack success rate increased to 5.8% (14/240), largely driven by grabbing (16.3%, 13/80). Pointing remained fully secure (0%), while typing resulted in a single false acceptance (1.3%). Notably, most successful attacks occurred when the attacker and victim had similar height, suggesting that physical similarity can make it easier to replicate the victim's movements.

Video-Replay Attack. This scenario yielded the highest attack success rate at 8.3% (20/240). Grabbing was again the most vulnerable task, with a 23.8% success rate. Pointing remained minimally affected (1.3%), and typing again showed complete robustness (0%). The increase from shoulder-surfing indicates that extended video exposure enables attackers to more effectively mimic grasping movements, particularly when they share similar physical traits with the victim.

Answer to RQ3: Typing and pointing remained highly resilient to observation-based attacks, with near-zero success rates even under video replay. In contrast, grabbing exhibited higher susceptibility, with success rates of 16.3% under shoulder-surfing and 23.8% under video replay. These results suggest that typing and pointing are preferable choices for secure behavioral authentication.

VII. Discussion

A. Task-Specific Trade-offs

Our evaluation shows that all three interaction tasks—grabbing, pointing, and typing—are viable for implicit authentication in VR, but each presents distinct trade-offs across accuracy, security, and usability, making them suitable for different deployment contexts.

Grabbing achieved a median EER of 0.4% in Scenario 2 (multi-session with SMOTE), offering strong accuracy with minimal user effort. It was rated "Easy" or "Very Easy" by 87.5% of participants, with 66.7% selecting "Very Easy," indicating excellent usability. However, grabbing was also the most vulnerable to observation-based attacks, with success rates of 16.3% under shoulder-surfing and 23.8% under video-replay. Its relatively uniform motion patterns across users may limit discriminability under adversarial settings. Future work may improve robustness by introducing randomized object placements or subtle gesture perturbations to diversify behavior without degrading usability.

Pointing showed high resilience to observation-based attacks, with no successful attempts under shoulder-surfing, and only 1.3% success rates under both no-knowledge and video-replay scenarios. It also received high usability ratings (87.5% "Easy" or "Very Easy"), suggesting broad user acceptance. However, its median EER remained higher than the other tasks (2.6% in Scenario 2), indicating relatively weaker discriminability. Incorporating more detailed behavioral cues, such as wrist movement, gesture timing, and gaze direction, can enhance accuracy while maintaining resistance to impersonation.

Typing achieved the lowest EER (0.3%) in Scenario 2 and showed exceptional resistance to all attack types, with only one successful impersonation observed (1.3% under video-replay). Its structured, ray-based input likely contributes to both high accuracy and user specificity. However, typing was perceived as the least usable task. Only 25% of participants rated it as "Easy" or "Very Easy," while 37.5% found it difficult. Usability enhancements such as predictive input, simplified keyboard layouts, or haptic feedback may help reduce fatigue and improve adoption.

In summary, *grabbing* offers high usability and low authentication error rates, making it a convenient option. However, due to its susceptibility to observation-based attacks, its use should be limited to trusted environments. *Pointing* strikes a good balance between usability and security, with low attack success rates and strong user acceptance. However, its lower accuracy compared to other tasks may cause usability issues, especially in scenarios where frequent false rejections can disrupt the user experience. When strong security is required, we recommend *typing* despite its lower usability, as it offers the highest accuracy and strongest resistance to impersonation.

B. Behavioral Consistency and Temporal Adaptation

Our results indicate that behavioral traits are not entirely static; instead, they evolve even within short timeframes. Models trained with single-session data (Scenario 1) showed noticeably higher EERs across all tasks, underscoring the challenges of generalizing from limited enrollment samples. In contrast, multi-session training (Scenario 2) improved accuracy substantially—especially for grabbing and typing—by capturing intra-user behavioral variability.

This finding suggests that practical VR authentication systems should incorporate temporal diversity in enrollment. At minimum, collecting behavioral samples from at least two sessions can help capture evolving interaction styles and reduce overfitting. Furthermore, authentication models should support adaptive updating or periodic re-enrollment to maintain reliability over time.

C. Practical Deployment Considerations

Our approach is designed for practical deployment in standalone VR headsets, emphasizing low-latency, privacy-preserving, and GPU-free operation. The lightweight SVM-based model achieves sub-second inference across all tasks (*e.g.*, 0.4—0.9 seconds) and operates entirely on-device, eliminating reliance on external servers or persistent connectivity. This is especially valuable in immersive applications, where network latency and privacy risks must be minimized.

One critical factor in real-world deployment is the management of reference user data. Since our system uses behavioral profiles from a small set of reference users to model the negative class, ensuring the quality and diversity of these samples is essential. These users must remain anonymous, and their data must be collected under appropriate consent and usage policies. In practice, reference sets should be curated to include behaviorally diverse, stable users and be periodically updated to reflect evolving population characteristics.

Initial deployment may suffer from limited data for the target user, particularly in short-term or cold-start scenarios. While our evaluation showed that SMOTE-based augmentation improves accuracy in multi-session settings, it offered little benefit under single-session conditions. This suggests that data diversity, not just quantity, is key to effective augmentation. Systems should therefore prioritize incremental data collection over time and consider adaptive learning strategies that gradually update user profiles without full retraining.

Additionally, fallback mechanisms can enhance reliability by initiating explicit re-authentication when implicit confidence drops below a certain threshold. Task-aware balancing between usability and security is also crucial. For example, while grabbing is more natural and usable, it is less robust to observation-based attacks. In contrast, typing offers stronger impersonation resistance, making it preferable in high-security contexts despite lower usability.

These considerations highlight that practical deployment involves more than just raw accuracy; it also requires sustainable learning, efficient on-device performance, and thoughtful integration into real-world usage.

D. Limitations

While our findings demonstrate the viability of implicit authentication via natural VR interactions, several limitations remain. First, our user study involved only 24 participants recruited from a single university community, limiting demographic diversity in age, culture, and interaction style. This homogeneity may restrict the generalizability of our results to broader populations. Future studies should include more diverse user samples to validate system robustness.

Second, our evaluations were conducted using a single VR device (Meta Quest Pro) in a controlled indoor environment. However, users in real-world settings may use diverse VR headsets (*e.g.*, HTC Vive, Valve Index, Sony PlayStation VR, or Pico) with varying tracking systems and hardware capabilities. Differences in tracking accuracy, sampling rates, and controller ergonomics can significantly impact the quality of behavioral features. These variations could influence behavioral features and authentication accuracy in real-world deployments, underscoring the need for cross-device evaluations to ensure robustness.

Lastly, our experiments focused on short-term behavioral drift over a few days. We did not investigate long-term drift spanning weeks or months, which may affect the system's ability to retain accuracy over time. As user behavior naturally evolves over time, maintaining model performance remains a challenge. Exploring adaptive learning mechanisms to address this temporal evolution remains an important direction for future work.

VIII. Conclusion

We presented a practical implicit authentication system for VR that uses natural interaction tasks—grabbing, pointing, and typing—as behavioral biometrics. Starting from 224 features extracted from head-mounted and controller sensors, we refined the model to 221 features based on ablation analysis, enabling the construction of a highly accurate authentication model. Our system achieves sub-second inference across all tasks, requires no GPU support, and consumes minimal CPU memory. This enables efficient on-device processing and real-time authentication, seamlessly integrating accurate, efficient, and secure implicit authentication into typical VR interactions.

In a 24-participant study, our method achieved strong performance, with median EERs of 0.4% for grabbing, 2.6% for pointing, and 0.3% for typing. Unlike many academic works that focus solely on accuracy without considering real-world requirements, our results demonstrate EERs that are not just low but sufficiently robust for practical authentication scenarios. Security evaluations confirmed that pointing and typing were highly resistant to observation-based attacks, while grabbing was significantly vulnerable under video replay, with a 23.8% success rate. Although typing proved ideal in terms of accuracy and security, participants found it less usable compared to grabbing and pointing.

As part of future work, we plan to conduct long-term, in-the-wild deployments to assess how authentication performance evolves over time and how users perceive the system in natural usage contexts. We aim to understand behavioral drift, session-to-session consistency, and the impact of implicit authentication on user comfort and trust.

Acknowledgment

We sincerely thank the reviewers for their valuable feedback. Hyoungshick Kim is the corresponding author. This work was supported by the IITP grant funded by MSIT (RS-2023-00229400, RS-2024-00436934, RS-2024-00459638, IITP-2025-RS-2022-II221199), and by the KISA grant funded by PIPC (No. 2780000017).

References

[1] F. J. Agbo, I. T. Sanusi, S. S. Oyelere, and J. Suhonen, “Application of Virtual Reality in Computer Science Education: A Systemic Review Based on Bibliometric and Content Analysis Methods,” *Education Sciences*, vol. 11, 2021.

[2] S. Karaosmanoglu, L. Kruse, S. Rings, and F. Steinicke, “Canoe VR: An Immersive Exergame to Support Cognitive and Physical Exercises of Older Adults,” in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA)*, 2022.

[3] M. Maciaś, A. Dąbrowski, J. Fraś, M. Karczewski, S. Puchalski, S. Tabaka, and P. Jaroszek, “Measuring Performance in Robotic Teleoperation Tasks with Virtual Reality Headgear,” in *Proceedings of the Progress in Automation, Robotics and Measurement Techniques (Automation)*, 2020.

[4] B. Odeleye, G. Loukas, R. Heartfield, G. Sakellari, E. Panaousis, and F. Spyridonis, “Virtually Secure: A Taxonomic Assessment of Cybersecurity Challenges in Virtual Reality Environments,” *Computers & Security*, vol. 124, 2023.

[5] L. Hallal, J. Rhinelander, and R. Venkat, “Recent Trends of Authentication Methods in Extended Reality: A Survey,” *Applied System Innovation*, vol. 7, 2024.

[6] 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 on the IEEE Symposium on Security and Privacy (SP)*, 2023.

[7] S. Stephenson, B. Pal, S. Fan, E. Fernandes, Y. Zhao, and R. Chatterjee, “SoK: Authentication in Augmented and Virtual Reality,” in *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, 2022.

[8] H. T. M. A. Riyadh, D. Bhardwaj, A. Dabrowski, and K. Krombholz, “Usable Authentication in Virtual Reality: Exploring the Usability of PINs and Gestures,” in *Proceedings of the International Conference on Applied Cryptography and Network Security (ACNS)*, 2024.

[9] A. Agarwal, R. Ramachandra, S. Venkatesh, and S. M. Prasanna, “Biometrics in Extended Reality: A Review,” *Discover Artificial Intelligence*, vol. 4, 2024.

[10] A. S B, A. Agrawal, Y. Yao, Y. Zou, and A. Das, “"What are they gonna do with my data?": Privacy Expectations, Concerns, and Behaviors in Virtual Reality,” in *Proceedings of the Privacy Enhancing Technologies (PETS)*, 2025.

[11] M. Li, N. K. Banerjee, and S. Banerjee, “Using Motion Forecasting for Behavior-based Virtual Reality (VR) Authentication,” in *Proceedings of the IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR)*, 2024.

[12] 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 (TDSC)*, vol. 16, 2018.

[13] X. Wang and Y. Zhang, “Nod to Auth: Fluent AR/VR Authentication with User Head-neck Modeling,” in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA)*, 2021.

[14] P. Arias-Cabarcos, T. Habrich, K. Becker, C. Becker, and T. Strufe, “Inexpensive Brainwave Authentication: New Techniques and Insights on User Acceptance,” in *Proceedings of the USENIX Security Symposium (USENIX Security)*, 2021.

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

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

[17] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *Journal of Artificial Iintelligence Research (JAIR)*, vol. 16, 2002.

[18] C. George, M. Khamis, E. von Zezschwitz, M. Burger, H. Schmidt, F. Alt, and H. Hussmann, “Seamless and Secure VR: Adapting and Evaluating Established Authentication Systems for Virtual Reality,” in *Proceedings of the Symposium on Usable Security and Privacy (USEC)*, 2017.

[19] 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 the Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2018.

[20] W. Shao, S. Luo, and Z. Yan, “Cross-content User Authentication in Virtual Reality,” in *Proceedings on Mobile Computing and Networking (MobiCom)*, 2024.

[21] D. Rupp, P. Grießer, A. Bonsch, and T. W. Kuhlen, “Authentication in Immersive Virtual Environments through Gesture-based Interaction with a Virtual Agent,” in *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 2024.

[22] J. Liebers, C. Burschik, U. Gruenefeld, and S. Schneegass, “Exploring the Stability of Behavioral Biometrics in Virtual Reality in a Remote Field Study: Towards Implicit and Continuous User Identification through Body Movements,” in *Proceedings of the ACM Symposium on Virtual Reality Software and Technology (VRST)*, 2023.

[23] R. Miller, N. K. Banerjee, and S. Banerjee, “Within-system and Cross-system Behavior-based Biometric Authentication in Virtual Reality,” in *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 2020.

[24] ——, “Using Siamese Neural Networks to Perform Cross-System Behavioral Authentication in Virtual Reality,” in *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 2021.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is All you Need,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[26] R. J. Teather and W. Stuerzlinger, “Pointing at 3D targets in a Stereo Head-tracked Virtual Environment,” in *Proceedings of the IEEE Symposium on 3D User Interfaces (3DUI)*, 2011.

[27] ISO, “9241-9 Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs)-Part 9: Requirements for Non-keyboard Input Devices (FDIS-Final Draft International Standard),” *International Organization for Standardization*, vol. 3, 2000.

[28] 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 the CHI Conference on Human Factors in Computing Systems (CHI)*, 2019.

[29] I. S. MacKenzie and R. W. Soukoreff, “Phrase Sets for Evaluating Text Entry Techniques,” in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA)*, 2003.

[30] W. Jeon, C. Lim, and H. Kim, “Exploring Natural Interactions for Implicit User Authentication in VR,” in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA)*, 2025.

[31] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[32] D. P. Kingma, “Auto-encoding Variational Bayes,” *arXiv preprint arXiv:1312.6114*, 2013.