Can Vibration Patterns Identify Users? Authentication for Smartphone-Watch Collaboration

Zicheng Cui, Zhihai Yang, Zhiquan He, Chenxu Kong, Jianhua He, Jianxin Li, Pinghui Wang, and Zhiquan Liu

***Abstract*—With increasing popularity of mobile smart devices user privacy and security are becoming particularly critical. Traditional authentication methods such as passwords and fingerprints, face the risk of forgery attacks and replay attacks, which are difficult to provide anti-imitation identity discrimination. This paper investigates the propagation characteristics of active vibration signals in hand-wrist coordination during natural touch operations by users, and models their response patterns. We found that there are differences in hand-wrist structure and micro-dynamic behavior among different individuals, which exhibits stable and distinguishable dynamic characteristics in vibration response. Based on the observation, we propose a dual-terminal joint identity authentication system (called *VIP*) for both smartphone and smartwatch. The system first generates vibration by dual-terminal motors and collects response signals using built in accelerometers, gyroscopes, and magnetometers to extract differential response patterns of users in natural touch. Then, we design a discriminative model, TouchFormer, that can extract direction sensitive features in multi-axis vibrations and achieve dynamic alignment and time compensation at critical moments, effectively integrating asynchronous and cross-device response information. Furthermore, we construct multi-device datasets of thirty users in real-world scenarios to validate the effectiveness of *VIP*. Extensive experiments demonstrate that *VIP* performs well on different devices, with an average authentication accuracy improvement of 5.65% compared to existing methods. In addition, the false acceptance rates of *VIP* under simulated attacks and replay attacks are 1.55% and 2.71%, respectively, which are on average 0.61 and 0.28 percentage points lower than existing methods.**

***Index Terms*—User authentication, vibration signal, dual-terminal collaboration.**

I. Introduction

THE widespread use of mobile smart devices has made their security issues increasingly prominent. According to the statistics, the number of smartphone users worldwide reached 4.88 billion by 2024, accounting for 60.42% of the world's total population [1]. Smartphones and wearable devices store a large amount of sensitive data for users, which raised widespread concerns about the security of smart devices. For example, there were 3,205 data breach incidents worldwide in 2023, affecting approximately 353 million people [2]. Therefore, preventing unauthorized user data leakage and personal privacy infringement is the primary security issue that need to be considered for mobile smart devices. Traditional authentication mechanisms use passwords [3], PINs [4], graphical passwords [5], Face IDs [6], fingerprints [7], to unlock mobile devices. However, these mechanisms are vulnerable to threats such as shoulder surfing attacks [8], password inference attacks [9], and replay attacks [10]. More importantly, traditional methods are difficult to provide a sustained identity discrimination mechanism in the event of identity information (e.g., password) being leaked.

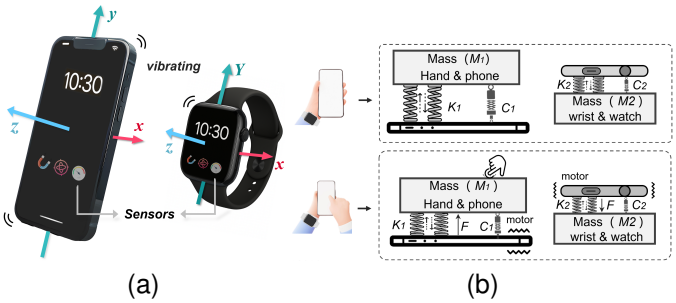


Fig. 1. Schematic diagram of active vibration excitation and response perception. (a) The coordinate system of the built-in sensors of the smartphone and smartwatch is used to capture multi-axis vibration response data. (b) Modeling of smartphone-watch collaborative vibration propagation.

To eliminate the above threats, a variety of identity authentication schemes for mobile devices have been proposed. A class of methods based on behavioral biometric utilize the user's operation behavior during natural interaction, such as walking gait [11], keystrokes [12], and touchscreen dynamics [13], to extract individual behavior patterns for identity authentication. However, such methods usually rely on specific interaction tasks and are vulnerable to simulation attacks. Although some studies have attempted to fuse multi-terminal sensor data to enhance authentication performance [14], there are still shortcomings in the collaborative modeling of cross-device data. Another major class of methods focuses on authentication mechanisms based on physiological biometrics, such as iris images [15], cardiac motion [16], hand geometry [17], and electrical signals caused by muscle contraction [18]. This type of method usually does not rely on continuous user interaction,

This work was supported in part by the National Natural Science Foundation of China under Grant 62172331 and in part by the Fundamental Research Funds for the Central Universities under Grant 300102404301, CHD.

Z. Cui, Z. Yang, Z. He, and C. Kong are with the School of Data Science and Artificial Intelligence, Chang'an University, Xi'an, China.

J. He is with the School of Computer Science and Electronic Engineering, University of Essex, UK.

J. Li is with the School of Business and Law, Edith Cowan University, Australia.

P. Wang is with the School of Cyber Science and Engineering, Xi'an Jiaotong University, Xi'an, China.

Z. Liu is with the College of Cyber Security, Jinan University, Guangzhou, China.

E-mail: {zichengcui, zhihaiyang, zhiquanhe, chenxukong}@chd.edu.cn; j.he@essex.ac.uk; jianxin.li@ecu.edu.au; phwang@mail.xjtu.edu.cn; zqliu@jnu.edu.cn.

but part of them relies on special hardware, such as electrodes, high-precision accelerometers, and electrocardiogram sensors, which is rarely configured in mainstream mobile devices. Therefore, there is still a lack of a solution that can be deployed on commercial mobile terminals and is practical and resistant to attacks.

In natural interactions with smart devices, the hands and wrists are the main contact and support parts. Their tissue structure (e.g., bones, muscles and skin density) and micro-dynamic behavior (e.g., pressure intensity and touch angle) jointly modulate the propagation and reflection of vibrations [19], [20], forming individual response patterns. Previous studies have shown that different hands [21] or wrist structures [22] have different responses to vibration excitation, resulting in specific response signals received by the accelerometer [23]. However, existing solutions only rely on a single hand or wrist area, making it difficult to reveal the propagation patterns between micro-dynamic behavioral disturbances and cross tissue structures. Additionally, a single sensor is difficult to capture the key response characteristics in the process of cross-device and cross-part transmission.

In this paper, We propose *VIP* (VIbration Pattern), a collaborative smartphone–smartwatch identity authentication system. *VIP* uses the vibration motors built into the smartphone and smartwatch to emit excitation signals, and combines inertial measurement units (IMUs), which include an accelerometer, gyroscope, and magnetometer, to collect responses in coordination. Ultimately, we construct a fine-grained representation of hand-wrist structural differences and behavioral differences under different axes and propagation paths. *VIP* can be regarded as a hybrid authentication mechanism that integrates micro-dynamic behavior and physiological structure attributes.To enhance the information dimension and discrimination accuracy, we collected multi-axis IMU data from dual-terminal devices.

It should be noted that there are still some challenges in achieving identity authentication for smartphone-watch collaboration. First, although it is known that the structure and micro-dynamic behavior of the hand-wrist affect the vibration response, it is still unclear how this effectiveness can be used for identity authentication. Second, different sensors and their different axes (see Figure 1(a)) have different sensitivities to vibration, which is still a challenge to fuse them for modeling. Third, a balance should be considered between system registration and system performance in order to provide a user-friendly experience.

To address the above challenges, we first verify the feasibility of using active vibration signals to capture hand-wrist response characteristics. Preliminary experiments show that different individuals have different hand-wrist vibration Patterns (HW-VIP). *VIP* generates vibration excitation from both the smartphone and the smartwatch through their built-in vibration motors (see Figure 1(b)), and synchronously collects response signals from the IMUs of both devices when the user touches the smartphone screen. To improve data quality, we filter the collected signals to suppress noise interference. Subsequently, we propose an encoder structure, TouchFormer, to address the fusion of dual-terminal vibration signals and to perform user authentication. TouchFormer combines two key modules: the dual-branch attention fusion block (D-BAFB) and the gated feedforward network (GFFN). Specifically, we use a dynamic-directional-attention (DDA) branch to model the sensitivity differences of different sensor axes in the vibration direction, and extract the micro-structural response characteristics of individuals in multi-axis directions. Furthermore, we also design a deformable-attention (DA) branch to build an interpolation mechanism based on time points and predicted offsets in order to enhance the system's alignment and compensation capabilities for critical moment responses. Additionally, GFFN can be used to enhance the nonlinear feature transformation capability by introducing a gating mechanism. To improve the generalization and robustness of the system in complex scenarios, moreover, we introduce a sharpness-aware minimization (SAM) optimization strategy [24] in model training. With the above design, the *VIP* can extract HW-VIP and can identify the user's identity through natural user interactions (such as clicking to unlock). Finally, *VIP* provides authorized access only to registered users.

In summary, the main contributions of this paper are:

- We propose a novel authentication system *VIP*, which utilizes active vibrations from both smartphone and smartwatch for identity authentication. *VIP* can be directly deployed on commercial mobile smart devices without the need for additional specialized hardware support.
- We investigate the feasibility of exploiting individual differences in hand–wrist structures and micro-dynamic behaviors for constructing HW-VIP, which originate from variations in physiological and behavioral characteristics among individuals, thereby revealing the potential of such factors for identity authentication.We use a dual-terminal IMU to synchronously collect response data for constructing user identity representation.
- We design a learning method *TouchFormer*, which uses D-BAFB as its core to construct direction sensitive features of vibration signals under multi axial, cross modal, and asynchronous conditions, ultimately enhancing the effectiveness of multi-source signal fusion and temporal modeling.
- We constructed user datasets with different genders, ages, and body structures, and designed a variety of natural interaction scenarios (e.g., wearing tightness, palm humidity) for actual measurement and evaluation. Experimental results demonstrate that the average authentication accuracy of *VIP* is significantly higher than that of competing benchmarks under different conditions. *VIP* also exhibits strong robustness in resisting simulation attacks and replay attacks.

II. BACKGROUND AND FEASIBILITY STUDY

A. Active Vibration Response Mechanism

As the main human-computer interaction medium, fingers have obvious individual differences in morphological structure and physiological parameters (e.g., size and bone density), which affect the overall response characteristics [25]. In this paper, we further construct two physical models in order to

analyze the excitation and response characteristics related to screen touching by fingers.

1) Vibration Generation: To describe the vibration transmission process in the hand-wrist structure, it is represented by a multi-degree-of-freedom viscoelastic system model, which consists of concentrated mass and series spring-damper units [26]. This structure can be abstracted into a one-dimensional or two-dimensional simplified propagation path model, where each node represents the equivalent mass of each tissue segment. Each connected element denotes the equivalent stiffness and energy dissipation characteristics of the soft tissue and joints. To simplify the analysis, it is assumed that the finger and the mobile phone maintain rigid contact (no slip) during the touch process.

As shown in Figure 1(b), mass M_1 represents the equivalent mass of the hand, and mass M_2 represents the equivalent mass of the wrist segment. The simple harmonic displacement excitation $y(t) = Y_0 \sin \omega t$ generated by the vibration motor acts on M_1 and M_2 through the soft tissue spring-damper unit (stiffness K_1, damping C_1). M_1 and M_2 are connected by another set of spring-damper units (stiffness K_2, damping C_2). Assuming that $x_1(t)$ and $x_2(t)$ represent the response displacements of the finger and wrist segment in the axial direction, respectively. According to Newton's second law, we can get:

$$\begin{cases} M_1\ddot{x}_1 + C_1(\dot{x}_1 - \dot{y}) + K_1(x_1 - y) \\ \qquad +C_2(\dot{x}_1 - \dot{x}_2) + K_2(x_1 - x_2) = 0 \\ M_2\ddot{x}_2 + C_2(\dot{x}_2 - \dot{x}_1) + K_2(x_2 - x_1) = 0 \end{cases} \tag{1}$$

where, $\dot{x}_i$ and $\ddot{x}_i$ represent the first and second derivatives respectively. The above model describes the complete path of vibration from the mobile phone to the hand and then propagates to the wrist through the joint structure. If the entire structure is simplified to a single degree of freedom system, its equivalent model can be simplified to a classic forced vibration second-order system:

$$m\ddot{x} + c\dot{x} + kx = F(t), \tag{2}$$

where $F(t)$ is the equivalent excitation force. Based on frequency domain analysis, the steady-state response of the system under the excitation force $F(t) = F_0 \sin \omega t$ is $x(t) = X_0 \sin(\omega t + \phi)$. Substituting into the equation and solving it, the following amplitude response can be obtained:

$$|X(\omega)| = \frac{F_0}{\sqrt{(k - m\omega^2)^2 + (c\omega)^2}}. \tag{3}$$

The phase lag is:

$$\phi(\omega) = \arctan\left(\frac{-c\omega}{k - m\omega^2}\right). \tag{4}$$

The natural frequency is denoted by w_n, which is calculated by $\omega_n = \sqrt{k/m}$. When the excitation frequency ω is close to ω_n, the system will resonate, resulting in a significant amplification of the response. The differences in m, c, and k among different individuals lead to differentiation of the frequency response characteristics in amplitude and phase. In addition, the difference in mechanical impedance of the contact interface further affects the transmission ratio of vibration energy between tissues:

$$g_1(t) = \frac{Z_1 - Z_2}{Z_1 + Z_2} f_1(t), \tag{5a}$$

$$f_2(t) = \frac{2Z_1}{Z_1 + Z_2} f_1(t), \tag{5b}$$

where Z_1 and Z_2 are the impedance at both terminals, $f_1(t)$, $f_2(t)$, and $g_1(t)$ denote the incident, transmitted, and reflected waves, respectively. Individual differences lead to different energy transmission ratios, thus forming a stable and discernible response pattern.

2) Nonlinear Vibration Response: Under external excitation, human soft tissues exhibit nonlinear stiffness and coupling behavior [27], making the system response no longer limited to the excitation frequency. On the contrary, this will be accompanied by frequency dimensionality increase. In other words, the emergence of higher-order harmonics and intermodulation frequency components significantly enhances the spectral complexity and user distinguishability of the response. The restoring force of soft tissue contains high-order nonlinear terms, and its mechanical behavior can be characterized by the Duffing equation:

$$m\ddot{x} + c\dot{x} + kx + \alpha x^3 = F_0 \cos(\omega t). \tag{6}$$

Under single-frequency excitation ω, we expand the steady-state response of the system into a harmonic series:

$$x(t) = \sum_{n=1}^{\infty} A_n \cos(n\omega t), \tag{7}$$

the amplitude A_n of each harmonic order is closely related to the nonlinear stiffness α of the tissue. Therefore, even under the same excitation, the nonlinear response spectrum distributions of different individuals are significantly different.

B. Feasibility Study

To verify the feasibility of authentication based on HW-VIP under active vibration excitation, we designed and implemented a series of experiments. All participants used a single handheld Xiaomi 15 smartphone and wore a Redmi Watch 5 smartwatch to complete data collection.

Based on the theoretical analysis in §II-A, it can be seen that the hand-wrist of different users, as a nonlinear propagation medium, produces individual modulation of the external vibration excitation signal, reflecting the user differentiation of HW-VIP. To enhance the observability of this feature in actual scenarios, our experiments rely on the built-in vibration motors of both the smartphone and the smartwatch to apply a fixed frequency excitation signal under stable conditions for verification. To minimize the interference of individual behavioral differences, all users perform touch tasks at the same position on the mobile phone screen, with a unified holding posture and wearing method.

In the specific experimental verification stage, we randomly selected two volunteers, collected their HW-VIP at different time periods, and extracted their frequency domain responses

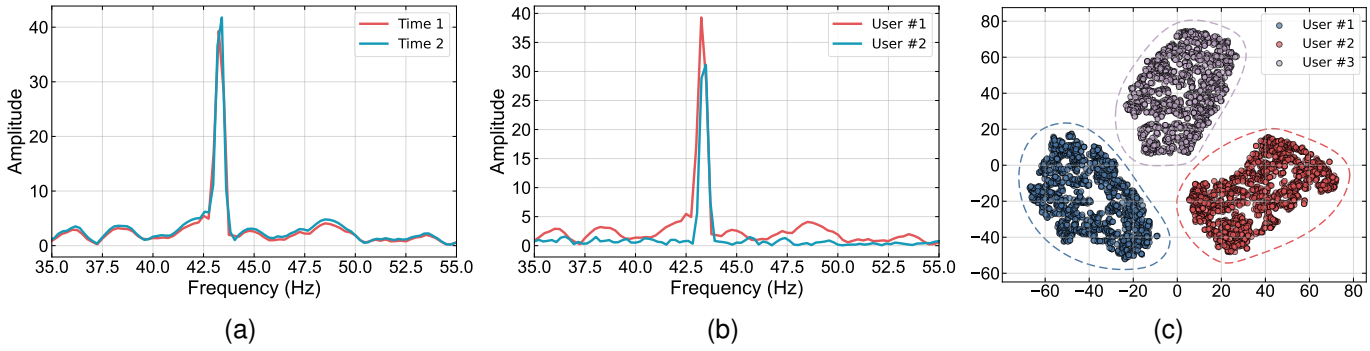


Fig. 2. Comparative analysis of individual HW-VIP under screen touch condition. (a) Frequency domain comparison of the same volunteer at two different times. (b) Frequency domain comparison of two different volunteers under the same conditions. (c) t-SNE visualization of HW-VIP from three different volunteers, demonstrating clear inter-user separability.

through fast fourier transform (FFT) [28]. Figure 2(a) shows the frequency spectrum distribution of the same volunteer at different time periods. The main frequency position and amplitude are highly consistent, which verifies the stability of the time domain response. Figure 2(b) reveals the differences in frequency domain characteristics of different volunteers, especially the main frequency amplitude. Further, we visualized the response characteristics of the three volunteers by t-SNE [29] dimensionality reduction. As shown in Figure 2(c), in the two-dimensional feature space, the vibration responses of the same volunteer are clustered in the same area and are clearly distinguished from other volunteers. In summary, the differences in the stable responses of different individuals to the same active excitation confirm the distinguishability of HW-VIP and provide theoretical basis and experimental support for *VIP*.

III. THREAT MODEL

In this paper we consider attacks that aim to bypass or deceive authentication system in order to obtain private information or perform unauthorized operations [30]. We assume that the attackers are unable to obtain physical access to the target device or intervene in the process of collecting registration data during the *VIP* training phase. Furthermore, we assume that the attackers possess the following capabilities:

- Attackers can obtain brief physical contact opportunities with the target device during user interaction gaps or when the device is idle.
- Attackers have obtained the identity information of the legitimate user and can use visual observation, video playback, or social engineering methods to obtain the target user's operational behavior (e.g., grip posture, touch method, finger position, and force pattern).
- Attackers have certain device deployment and signal reconstruction capabilities, and can use universal sensor modules (i.e., external IMU) to complete vibration acquisition, processing, and synthesis operations, and finally convert them into control signals that can be used for reproducing target vibrations in LRAs.

To this end, we consider the following attack methods:

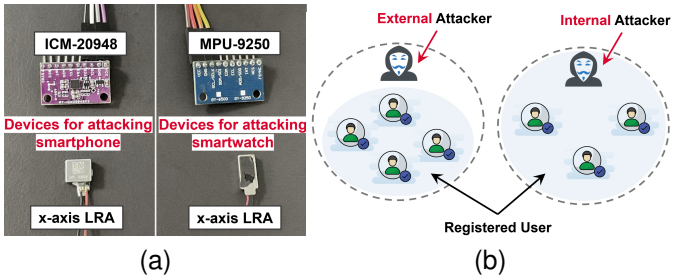


Fig. 3. Threat model diagram. (a) Replay attack via Side-channel sensing. (b) Impersonation attack.

1) Side-channel Replay Attack: The attacker secretly arranges a high-precision external IMU (see Figure 3(a)) near the device when the user is using it (assuming the user is not aware of it), collecting vibration response signals triggered by the user's interaction with the device. Subsequently, the attacker will replay the collected sensor data to the target device through an external LRA (see Figure 3(a)), attempting to induce its sensors to receive HW-VIP consistent with legitimate users, thereby bypassing the *VIP* system [31].

2) Impersonation Attack: The attacker attempts to simulate the victim's tactile behavior by directly contacting the target device, in order to deceive the authentication system [30]. Attackers may improve the success rate of attacks by selecting individuals with similar hand structures (such as size and mass distribution), or by observing and imitating the interaction patterns of the victims (e.g., touch position, force, and grip). In addition, considering that the system supports multi-user registration, we further divide impersonation attacks into two scenarios (see Figure 3(b)):

- *Internal impersonation attacks*: The attacker is a registered user within the system, but he attempts to impersonate the identities of other registered users in the system for authentication.
- *External impersonation attack*: The attacker is not registered in the system and attempts to illegally gain access by imitating the interaction behavior of legitimate users.

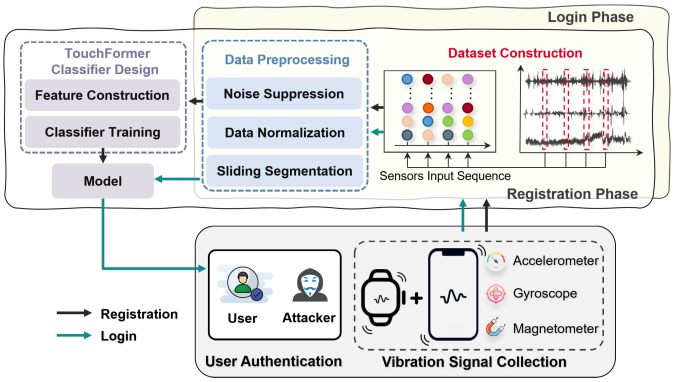


Fig. 4. System architecture of *VIP*.

IV. SYSTEM DESIGN

A. Overview

The *VIP* system architecture is shown in Figure 4. When the user touches the screen, the dual-terminal devices generate active excitation signals simultaneously through the vibration motor. This signal is affected by the physical structure and micro-behavior of the user's finger-wrist, and the built-in IMU (i.e., accelerometer, gyroscope, and magnetometer) of the dual-terminal devices synchronously collect response data.

During the registration phase, the system first preprocesses the collected data, including denoising, normalization, and segmentation. Specifically, *VIP* uses a bandpass filter to extract the response signal in the main frequency band, suppressing low-frequency drift and high-frequency noise caused by hand movements and environmental disturbances. Subsequently, *VIP* normalizes the three-axis IMU data, and introduces a sliding window mechanism to extract multiple locally aligned response segments in the time domain. Finally, they are input into the subsequent authentication model and stored for the login stage.

To further characterize the dynamic response characteristics, we construct a Transformer-based identity authentication model called *TouchFormer*. The core of *TouchFormer* includes D-BAFB and GFFN modules, which we will introduce in detail in §IV-E and §IV-F, respectively. In addition, in the login phase, *VIP* first captures the vibration signal generated by the dual-terminal device motors when a finger touches the phone, and then performs data preprocessing in the same way as the registration phase. Then, *VIP* uses the model trained in registration phase to perform user identity authentication and deception detection.

B. Vibration Signal Collection and Dataset Construction

Most smartphones and smartwatches have built-in Linear Resonant Actuator (LRA), with typical operating frequencies generally distributed within 300 Hz, and controllable vibration amplitude. In this work, we use smart devices with LRAs as excitation sources and simultaneously collect response data via their built-in IMUs during touch-vibration.

Specifically, during a period of touch-vibration duration t, the IMU continuously records sensor data at a sampling frequency f. The number of raw data points collected by each single-axis sensor s ($s = a, g, m$, respectively represents the accelerometer, gyroscope, and magnetometer) during this period is $n = t \times f$. Taking the gyroscope as an example, its three-axis angular velocity data can be expressed as a $3 \times n$ matrix :$D_g = [\mathbf{g}_1, \mathbf{g}_2, \cdots, \mathbf{g}_n]$, $\mathbf{g}_i = \begin{bmatrix} g_x^i & g_y^i & g_z^i \end{bmatrix}^\top$. Similarly, the raw data of the accelerometer and magnetometer are recorded as D_a and D_m, respectively. For dual-terminal devices e ($e = p, w$, where $e = p, w$ indicates the smartphone and smartwatch, respectively), their collected multimodal sensor data can be uniformly expressed as: $D_e = [D_{a,e}, D_{g,e}, D_{m,e}]$, where D_e is a $9 \times n$-dimensional composite sensor data matrix, $D_{a,e}, D_{g,e}, D_{m,e}$ correspond to the data sub-matrices of the accelerometer, gyroscope, and magnetometer on the device e, respectively.

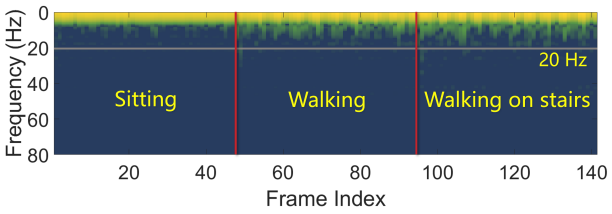


Fig. 5. Frequency distribution of noise caused by human motion in three common scenarios.

C. Data Preprocessing

1) Motion and Environmental Noise Suppression: The vibration signal contains not only effective responses from touch, but also low-frequency noise caused by the user's overall movements (e.g., walking and wrist swinging). Experimental statistics show that the biometric response caused by touch vibration operation is mainly distributed above 30 Hz, while the noise caused by human body movement is mainly distributed below 20 Hz, as shown in Figure 5. To resist low-frequency and high-frequency noise interference, we uses a bandpass filter [32] to process the raw data $d_{s,i}(k)$ of each triaxial sensor s:

$$\hat{d}_{s,i}(k) = \mathcal{F}(d_{s,i}(k)), \tag{8}$$

where $\mathcal{F}(\cdot)$ represents bandpass filtering.

2) Data Normalization and Sliding Segmentation: To unify the numerical range of different sensor channels, the bandpass filtering result $\hat{d}_{s,i}(k)$ of each sensor axis $i \in \{x, y, z\}$ is normalized as:

$$\tilde{d}_{s,i}(k) = \frac{\hat{d}_{s,i}(k) - \min_k \hat{d}_{s,i}(k)}{\max_k \hat{d}_{s,i}(k) - \min_k \hat{d}_{s,i}(k)}, \tag{9}$$

where $\min_k \hat{d}_{s,i}(k)$ and $\max_k \hat{d}_{s,i}(k)$ represent the minimum and maximum values of the sampling sequence of sensor s on this axis, respectively, and $\tilde{d}_{s,i}(k)$ represents the normalized result. The normalized multi-sensor data on device e is uniformly represented as: $\tilde{D}_e = [\tilde{D}_{a,e}, \tilde{D}_{g,e}, \tilde{D}_{m,e}]$.

Then, we use a sliding window with a length of $w = 200$ and a step size of $s = 100$ for slicing to ensure that each segment covers a complete touch event. For the normalized

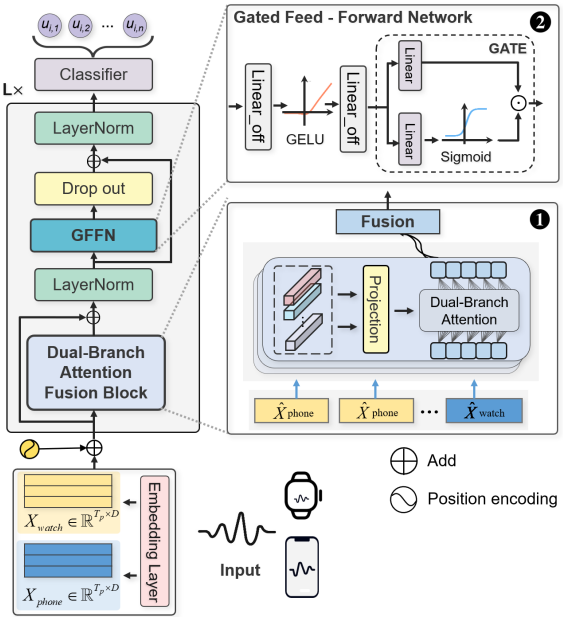


Fig. 6. Architecture of *TouchFormer*, which consists of a Dual-Branch Attention Fusion Block (❶) and a Gated Feed-Forward Network (❷) as its core components.

data $\tilde{D}_e \in \mathbb{R}^{9\times n}$ of each device e, we finally get the slice sequence: $\mathcal{W}_e = \left\{\tilde{D}_e^{(1)}, \tilde{D}_e^{(2)}, \ldots, \tilde{D}_e^{(\bar{N})}\right\}$, $\tilde{D}_e^{(j)} \in \mathbb{R}^{9\times w}$, where the j-th slice represents the j-th segment of sensor response data after normalization.

D. Collaborative Vibration Feature Modeling

To effectively capture the multi-axis vibration characteristics of the hand-wrist, as shown in Figure 6, the input of *TouchFormer* is the nine-axis IMU sensor data from the mobile phone and watch, denoted as $D_p = [D_{a,p}, D_{g,p}, D_{m,p}] \in \mathbb{R}^{T_p\times d}$ and $D_w = [D_{a,w}, D_{g,w}, D_{m,w}] \in \mathbb{R}^{T_w\times d}$ respectively, where D_a, D_g, D_m represent the $x/y/z$ three-axis data of the accelerometer, gyroscope, and magnetometer respectively (see Figure 1(a)), and the feature dimension d is 9. We use two independent embedding layers $E_p(\cdot)$, $E_w(\cdot)$ to map D_p and D_w to a unified dimensional space: $X_p = E_p(D_p) \in \mathbb{R}^{T_p\times D}$, $X_w = E_w(D_w) \in \mathbb{R}^{T_p\times D}$, and add a learnable position encoding P to retain the temporal information: $\hat{X}_p = X_p + P_p$, $\hat{X}_w = X_w + P_w$. Subsequently, the fused input $X = \left[\hat{X}_p, \hat{X}_w\right]$ is sent to the *TouchFormer* composed of L layers of encoders stacked together (see Figure 6). Finally, the output of *TouchFormer* is used for identity classification tasks.

E. Dual-Branch Attention Fusion Block

1) Dynamic Directional Attention (DDA): To more effectively model the sensitivity differences between dual-terminal devices in different directions, inspired by SDformer [33], we proposed an improved DDA module. As shown in Figure 7, this module is designed to model the dynamic response characteristics in cross-device transmission, especially short-term high-frequency disturbances (such as vibration spikes) that are easily ignored by traditional attention mechanisms.

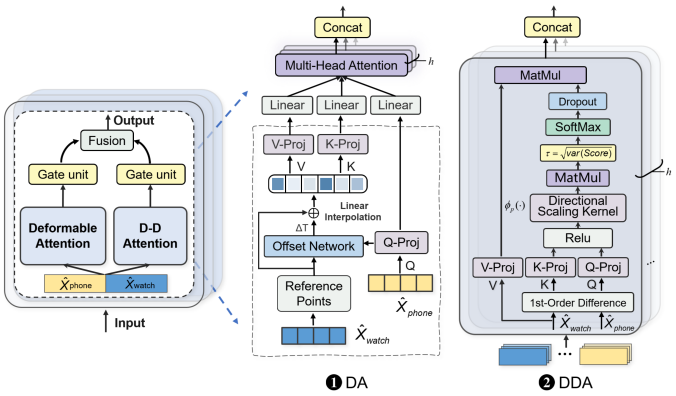


Fig. 7. The architecture of Dual-Branch Attention Fusion Block, the core of which includes DA (❶) and DDA (❷).

First, we add a first-order difference operator $\nabla x_t = x_t - x_{t-1}$ to the input IMU sequence x_t before Q^{dda} and K^{dda} projection, which is used to calculate the local temporal mutation of the sequence to guide the attention module to focus on the starting point of the change in the signal before $Q^{\text{dda}} \in \mathbb{R}^{L\times d}$, $K^{\text{dda}} \in \mathbb{R}^{L\times d}$ projection, such as sudden touch vibration.

After that, we introduce a nonlinear kernel function $\phi_p(\cdot)$ for each Q^{dda} and K^{dda} vector, which is defined as follows:

$$\phi_p(x) = f_p(\text{ReLU}(x)), \tag{10a}$$

$$f_p(x) = x \odot W_{\text{dir}} \odot (\text{std}(x))^{-p} \cdot \lambda_{\text{dyn}}, \tag{10b}$$

where x is the input vector, W_{dir} is the learnable directional weight vector, $\text{std}(x)$ represents standard deviation normalization, and λ_{dyn} is a learnable dynamic scaling factor. Specifically, λ_{dyn} enables the model to adaptively adjust the attention weight according to the local perturbation strength. λ_{dyn} can enhance the area when a strong mutation (such as sudden acceleration) is detected. Additionally, λ_{dyn} appropriately reduces the attention to the signal in a smooth trend change area (such as hand-steady touch). The power exponent p redirects the Q^{dda}, K^{dda} through std normalization. A larger p helps to strongly cluster the local directional patterns, but may suppress certain weak but directionally significant feature components. Experimental results demonstrate that a proper p balances transient response capture and trend stability.

After the transformation of ϕ_p, Q^{dda} and K^{dda} are recorded as $Q'^{\text{dda}} = \phi_p\left(Q^{\text{dda}}\right)$ and $K'^{\text{dda}} = \phi_p\left(K^{\text{dda}}\right)$ respectively. The attention score is calculated:

$$\text{Score}_{ij}^{\text{dda}}(Q_i, K_j) = \phi_p\left(Q_i'^{\text{dda}}\right)\phi_p\left(K_{\text{j}}'^{\text{dda}}\right)^T. \tag{11}$$

To deal with the weight offset caused by the difference in device amplitude, we introduce a dynamic scaling factor τ, which is adaptively adjusted according to the variance of the current score matrix: $\tau = \sqrt{\text{Var}\left(\text{Score}^{\text{dda}}\right)}$. It adjusts the signal amplitude together with λ_{dyn} to alleviate attention bias caused by device differences. Further, the attention weight is calculated through Softmax and Dropout:

$$A_{ij}^{\text{dda}} = \text{Dropout}\left(\frac{\exp\left(\text{Score}_{ij}^{\text{dda}}/\tau\right)}{\sum_k \exp\left(\text{Score}_{ik}^{\text{dda}}/\tau\right)}\right). \tag{12}$$

The final output is $\text{Output}^{\text{dda}} = \sum_j A_{ij}^{\text{dda}} V_j^{\text{dda}}$, where V^{dda} is the value vector.

2) Deformable Attention (DA): To adaptively mine discriminative features from active touch vibration signals, we achieve cross-time domain key response modeling by learning sampling time offset, as shown in Figure 7. Specifically, within a window with a sampling sequence length of P, let the p-th reference point be t^p, the attention module predicts a time offset $\Delta t^p = \mathcal{F}_\Delta \left(Q^{\text{def}}\right)_p$ through the coordinate offset network $\mathcal{F}_\Delta$ [34]. To get the target sampling time of the current query Q^{def}, we have: $T_{\text{samp}}^p = t^p + \Delta t^p$. Since T_{samp}^p is generally a non-integer time point, we use the interpolation function $\mathcal{I}(\cdot)$ to obtain the corresponding features on the original signal X:

$$X\left(T_{\text{samp}}^p\right) = \mathcal{I}\left(X, T_{\text{samp}}^p\right), \tag{13}$$

this ensures that even if the target sampling point is not aligned with the sensor timing grid, effective information can be smoothly extracted. This interpolation mechanism is used to fuse the signals from X_P and X_W to alleviate the alignment error caused by the difference in sampling frequency of the IMU module.

Further, we linearly map T_{samp}^p to the attention space to obtain the key and value vectors: $K^p = W_{\text{def}}^K X\left(T_{\text{samp}}^p\right)$,$V^p = W_{\text{def}}^V X\left(T_{\text{samp}}^p\right)$. Among them, K^p and V^p represent the attention key and value corresponding to the sampling point, and the representative local features can be extracted through mapping. The dot product of Q^{def} and K^p is then calculated to obtain the attention score, the corresponding V^p is weighted summed, and $\sqrt{d}$ represents the scaling factor:

$$\begin{aligned} Z_{\text{head}}^{(h)} &= \text{Softmax}\left(\frac{Q^{def} K^{pT}}{\sqrt{d}}\right) V^P, \\ \text{s.t.} &\begin{cases} Q^{def} & \leftarrow \hat{X}_{\text{phone}} \\ K^p, V^p & \leftarrow \text{Offset} + \text{Interp}\left(\hat{X}_{\text{watch}}\right). \end{cases} \end{aligned} \tag{14}$$

The final Multi-Head Attention (M-HA) [35] output is:

$$\begin{aligned} \text{Output}^{\text{def}} &= \text{MHA}\left(Q^{\text{def}}, K^{\text{P}}, V^{\text{P}}\right) \\ &= \text{Concat}\left(\mathbf{Z}_{\text{head}}^{(1)}, \ldots, \mathbf{Z}_{\text{head}}^{(H)}\right) W^O. \end{aligned} \tag{15}$$

In order to further integrate the direction perception features and frequency mutation information, we input the outputs of DDA and DA into the gate unit for weighted control:

$$\hat{O}^{dda} = G^{dda} \odot \text{Output}^{dda}, \hat{O}^{def} = G^{def} \odot \text{Output}^{def}, \tag{16}$$

where G^{dda} , G^{def} are the learnable gate weight, $\odot$ represents element-by-element multiplication. The fused output is:

$$O^{D-BAFB} = W\,\text{Concat}\left(\hat{O}^{dda}, \hat{O}^{def}\right) + b, \tag{17}$$

where W and b are learnable parameters.

F. Gated Feed-Forward Network

The module adopts the Pre-Norm residual structure [35], where Layer Normalization (LN) is applied, and the output is connected with a residual connection: $U = LN\left(X + O^{D-BAFB}\right)$. It is then fed into GFFN to perform nonlinear transformation and gated fusion on the attention fusion features [36]. Its structure is shown in Figure 6 and can be divided into the following two levels:

1) Linear Transformation + GELU Activation: The input U is first linearly expanded, $GELU$ activated and linearly mapped to the original dimension. Given the input $\widetilde{U}^{(l)}$ of the l-th layer coding block normalized by the previous layer, this process can be expressed as $R^{(l)} = W_2^{(l)}\left(\text{GELU}\left(W_1^{(l)}\widetilde{U}^{(l)}\right)\right)$, where $W_1^{(l)}$ and $W_2^{(l)}$ are the weight matrices of the two-layer linear mapping.

2) The GLU Gating Mechanism: Generates gating signals and candidate information through two parallel linear branches: one branch obtains $G_a^{(l)} = W_3^{(l)} R^{(l)}$ through linear mapping, and the other branch obtains the gating vector $\sigma\left(G_b^{(l)}\right) = \sigma\left(W_4^{(l)} R^{(l)}\right)$ through Sigmoid activation after linear mapping. The gated output is:

$$W^{(l)} = G_a^{(l)} \odot \sigma\left(G_b^{(l)}\right). \tag{18}$$

Then, after adding it to the residual U, the output of the encoding block is:

$$H^{(l)} = LN\left(U^{(l)} + \text{Dropout}\left(W^{(l)}\right)\right). \tag{19}$$

Fig. 8. The prototypes of *VIP*.

V. Implementation and Evaluation

In this section we implement the *VIP* system and evaluate its performance. We will focus on answering the following six questions: (Q1) Does *VIP* have accuracy and cross-device adaptability? (Q2) Can *VIP* work stably under different vibration settings and external environments? (Q3) Does *VIP* adapt to different user behaviors and operating modes? (Q4) Is *TouchFormer* irreplaceable? (Q5) Can *VIP* effectively resist different types of attacks? and (Q6) Does *VIP* have long-term stability?

A. Experimental Settings

This section describes the experimental setup, including dataset collection, model training, and evaluation metrics.

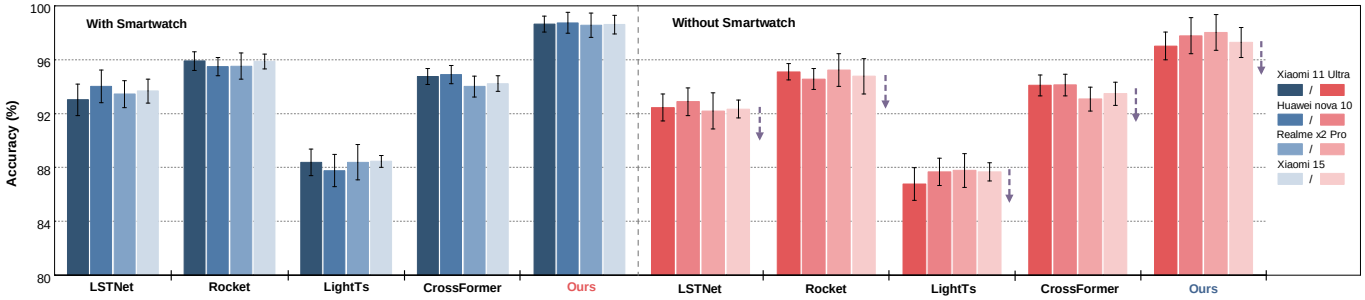


Fig. 9. Overall authentication accuracy of the five authentication models.

1) Data collection: To collect experimental data, we developed an Android-based data collection tool (see Figure 8) deployed on four models of smartphones (Xiaomi 11 Ultra, Huawei nova 10, Realme X2 Pro, and Xiaomi 15) and Redmi Watch 5, which is based on a common API interface to ensure cross-device consistency. We recruited a total of 41 volunteers aged between 20 and 56, and excluded 4 individuals with significant noise and partially missing data. Ultimately, 30 legitimate users (including 14 females and 16 males, numbered $U = \{U_1, U_2, \ldots, U_{30}\}$) and 7 external impersonation attackers (including 3 females and 4 males, numbered $U = \{U_{31}, U_{32}, \ldots, U_{37}\}$) will be retained. The entire experiment lasted for four months, and all procedures were approved by our Institutional Review Board (IRB).

To ensure data quality, we guide volunteers to familiarize themselves with the equipment and app operation procedures before formally collecting data, ensuring that they complete experimental tasks in their natural state. The specific experimental steps are as follows: 1) Volunteers need to hold a mobile phone and wear a watch at the same time, and are told to choose the appropriate handheld holding angle and wearing posture according to their own habits; 2) Volunteers are required to complete five consecutive intelligent touch screen interaction tasks in a common grip method (one handed or two handed) to collect individual touch vibration characteristics with posture differences; 3) Sampling was conducted under different experimental conditions (including sitting, walking, and walking on stairs), with 20 sets of touch vibration samples completed daily. To reduce the interference of task fatigue on data quality, we have controlled the interaction interval and prompt method; and 4) We synchronize the readings of the IMU sensor built in the recording device and store them in CSV format, with a sampling rate of $f = 100Hz$. For dual-terminal devices, each user's sensor data can be represented as $D_e = [D_{a,e}, D_{g,e}, D_{m,e}]$ for training *TouchFormer*.

2) Training Settings: During the registration phase, we use data from 30 legitimate users ($U = \{U_1, U_2, \ldots, U_{30}\}$) to train *TouchFormer* with a batch size of 256. Specifically, we use a 5-fold cross validation strategy to divide all samples into five equal parts after shuffling at the sample level, so that each fold preserves the sample distribution of each category as much as possible. In each round, one part is used as the test set $\mathcal{T}_e^{\text{test}}$, and the remaining four parts are used as the training set $\mathcal{T}_e^{\text{train}}$. 20% of the data is drawn from it to construct the validation set $\mathcal{T}_e^{\text{val}}$. To evaluate model generalization, we conduct m experiments ($5 < m < 10$) per iteration and report the average as the final metric.

During the training of *TouchFormer*, we used a cross-entropy loss function and combined it with a SAM optimizer to perform 200 rounds of iterative training on the model. The initial learning rate is 0.000001, and the weight decay is 0.00001. SAM introduces adversarial perturbation direction during each gradient update to optimize the model in a flatter region of the parameter space, where the perturbation radius ρ is set to 0.7.

3) Metrics: We use the false acceptance rate (FAR), false rejection rate (FRR), and F1-score as the main evaluation metrics. Among them: $FAR = \frac{FP}{FP+TN}$, represents the probability of misidentifying an imposter as a legitimate user; $FRR = \frac{FN}{FN+TP}$, stands for the probability of misidentifying legitimate users as imposters; and F1-score $= \frac{2\times \text{ Precision } \times \text{ Recall}}{\text{Precision } + \text{ Recall}}$, reflects the comprehensive performance of the system.

B. Overall Performance (Q1)

We compared the performance of *TouchFormer* with four candidate classification models (LSTNet [37], Rocket [38], LightTs [39], and CrossFormer [40]). All models were trained under a unified data preprocessing process as mentioned in §IV-C, following the same training method as introduced in §V-A. Their authentication performance was evaluated on the test set $\mathcal{T}_e^{\text{test}}$. In addition, to verify the transferability of the model on heterogeneous hardware platforms, we conducted experiments on four different models of smartphones.

Figure 9 illustrates the average accuracy of each model on a dataset of 30 legitimate users. The results demonstrate that *TouchFormer* performs the best in almost all settings, with an average accuracy of 98.65%, which significantly outperforms baseline methods. Each classification model exhibits consistent performance across different devices, indicating that the vibration features extracted by *VIP* have good cross-device adaptability. In addition, Rocket achieves suboptimal performance in the experiments with an average accuracy of 95.9%. It extracts discriminative features through multiple sets of random convolution kernels, demonstrating good generalization ability and computational efficiency. However, due to the lack of targeted structural awareness design, it is difficult to robustly model key timing dependencies in multi pose and

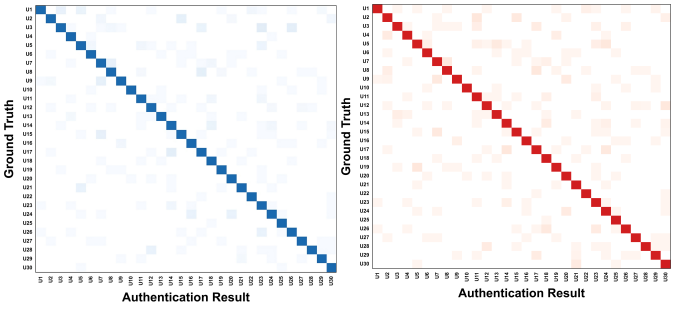


Fig. 10. Confusion matrix of *TouchFormer* on 30 users. (a) With smartwatch. (b) Without smartwatch.

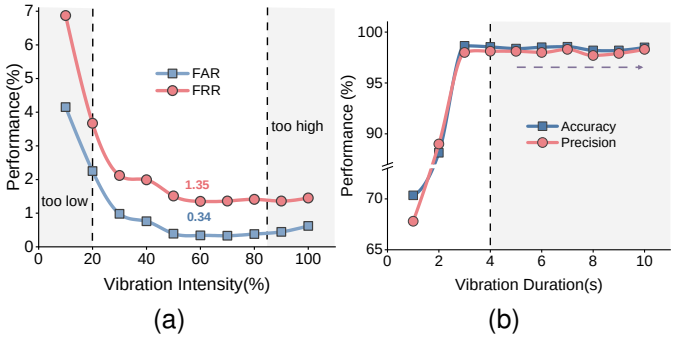


Fig. 11. Impact of vibration intensity and duration.

heterogeneous devices, which limits its performance. Furthermore, we plot a confusion matrix of 30 users obtained using the Xiaomi 15 as shown in Figure 10(a). It can be observed that the system has clear boundaries for distinguishing between different users, reflecting its sensitivity to identity features.

Further, to investigate the impact of watches on authentication performance, we excluded watch data and retrained the model. The right side of Figure 9 demonstrates that, in all device tests, the overall accuracy of users decreased by about 0% - 3%. The confusion matrix of 30 users (see Figure 10(b)) further indicates that the absence of wrist vibration features can lead to a certain decrease in *VIP* authentication performance. Compared with capturing hand vibration response solely through smartphone, integrating wrist features captured by smartwatch provides a complete HW-VIP composed of hand-wrist response features, which can more comprehensively capture users' physical structural characteristics and micro-dynamic behaviors.

C. Impact of Vibration Parameters (Q2)

To systematically evaluate the authentication performance of *VIP* under different vibration parameter configurations, we designed and implemented two sets of experiments focusing on the two key factors of vibration intensity and vibration duration, and analyzed their comprehensive impact on authentication accuracy and user experience.

***1) Impact of Vibration Intensities*:** As *VIP* relies on active vibration signals for identity authentication, we first examined the impact of different vibration intensities on authentication performance. The experiment was conducted in collaboration between dual-terminal devices (i.e., smartphone and smartwatch) using consistent vibration intensity. As shown in Figure 11(a), when the vibration intensity is set to 60%, the FAR and FRR of the system reach their optimal values of 0.34% and 1.35%, respectively. Under this intensity configuration, the system can stimulate clearer HW-VIP, thereby further improving feature separability.

When the vibration intensity is low (e.g., less than 20%), the extracted HW-VIP is close to the sensor background noise and is susceptible to interference due to the weak signal amplitude, resulting in decreased stability. When the vibration intensity is too high (e.g., greater than 85%), severe mechanical stimulation may interfere with the user's natural grip posture. Especially when wearing a watch, it is more likely to induce a "loosening effect" that leads to feature deviation, thereby disrupting the stability of HW-VIP. The Table I shows that most users are more inclined to accept moderate vibration intensity (about 60%) as the default con-

TABLE I
ANALYSIS OF SUBJECTIVE FEEDBACK FROM USERS ON VIBRATION SETTINGS. HERE, ○, ◑, AND ● REPRESENT ACCEPTABLE, BARELY ACCEPTABLE, AND UNACCEPTABLE RESPECTIVELY.

User ID	Vibration Intensity			Vibration Duration		
	$\leq 20\%$	$\approx 60\%$	$\geq 85\%$	< 4 s	4–7 s	> 7 s
U1	○	○	◑	○	○	◑
U2	○	○	○	○	◑	●
U3	○	○	○	○	◑	◑
U4	○	○	●	○	◑	◑
U5	○	○	○	○	○	◑
U6	○	○	●	○	◑	●
U7	○	○	◑	○	○	◑
U8	○	○	◑	○	○	◑
U9	○	○	○	○	◑	●
U10	○	○	○	○	◑	◑
U11	○	○	●	○	◑	●
U12	○	○	◑	○	◑	◑
U13	○	○	●	○	○	◑
U14	○	○	◑	○	◑	◑
U15	○	○	○	○	◑	●
U16	◑	◑	◑	○	○	◑
U17	○	○	○	○	○	◑
U18	○	○	◑	○	◑	●
U19	○	○	◑	○	◑	◑
U20	○	○	○	○	○	●
U21	○	○	◑	○	◑	◑
U22	○	○	◑	○	○	●
U23	○	○	◑	○	◑	◑
U24	○	○	◑	○	◑	●
U25	○	○	○	○	○	◑
U26	○	○	◑	○	◑	◑
U27	○	○	◑	○	◑	●
U28	○	○	●	○	◑	◑
U29	○	○	◑	◑	◑	●
U30	◑	◑	●	○	◑	◑

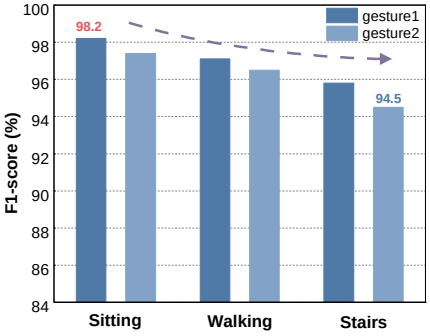


Fig. 12. F1-score under different activity scenarios and holding gestures.

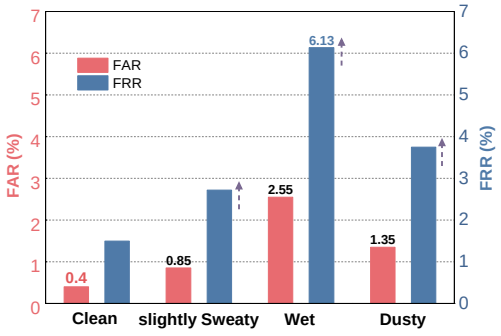


Fig. 13. Authentication performance under different finger conditions.

figuration for the authentication process. More than 90% of users expressed "completely acceptable" for a strength of approximately equal to 60%, which is highly consistent with the system's stable extraction capability for HW-VIP response signals under this configuration.When the vibration intensity is too low (e.g., less than or equal to 20%), the amplitude of the excited skin response signal is insufficient, resulting in lower energy and decreased stability of the IMU signal collected by the system. Subjectively, users provide feedback on "blurry vibration sensation" and "as if there is no vibration", which objectively manifests as a decrease in signal-to-noise ratio and repeatability, and has a negative impact on authentication performance.

On the contrary, when the intensity reaches greater than or equal to 85%, while the signal excited by the system is enhanced, the user experience also significantly deteriorates. In the survey, multiple users expressed dissatisfaction using words such as "harsh", "irritating", and "overly intense", and even received feedback such as "numbness in the hands" or "unstable grip on the phone". When wearing the watch, users mentioned that strong earthquakes are directly transmitted to the wrist through the strap, causing muscle tension or short-term tremors, resulting in a negative impression of "anxiety". More importantly, excessive intensity did not improve *VIP* authentication performance, but instead undermined users' trust and interaction habits. Based on a comprehensive analysis of authentication performance and user experience, we ultimately set the default vibration intensity of the device to 60%. According to user surveys, this configuration does not interfere with daily use while maintaining the user experience.

2) Impact of Touch-vibration Durations: We further investigated the impact of vibration duration on authentication performance. During the registration phase, the length of collected samples increases accordingly as the duration of vibration increases, which helps improve the training effectiveness and classification ability of the model. However, excessive duration will increase the user's interaction burden, affecting the system's response efficiency and actual experience. Figure 11(b) presents the variation curves of model accuracy and precision under different vibration durations.

The results indicate that the performance in the initial stage rapidly improves as the vibration duration increases. When the vibration duration reaches the third seconds, the system accuracy and precision reach 98.65% and 98.12%, respectively. After the vibration duration reaches the 4th second, the performance tends to stabilize and the improvement in performance slows down.

As shown in Table I, almost all users indicate that a single vibration duration of less than 4 seconds is "completely acceptable", which is perceived to be closer to the natural rhythm of interaction. When the duration was extended to more than 7 seconds, a significant proportion of users expressed "inability to accept", and some even gave feedback that "the time was too long and they lacked patience" and "they thought the system had a problem", especially among users in the middle and later stages. This reaction not only reduces the user's tolerance, but may also affect their perception of system reliability.

This trend is also validated in Figure 11(b), where the system reaches the optimal balance between performance and experience at 3-4 seconds, with an accuracy rate of 98.65%. Afterwards, the performance improvement tended to saturate, while the user experience rapidly declined. Therefore, we suggest setting the duration of each touch during the registration phase to 3.5 seconds. This not only fully ensures the richness of response sampling, but also effectively controls the perceived burden on users, ensuring the acceptability and practicality of the system in actual deployment.

D. Performance under Varying Environments (Q2)

In practical usage environments, users may be in various daily activity states, such as sitting still, walking, or walking on stairs, which can cause background noise, hand tremors, and posture changes that may interfere with the stability of the original tactile signal. At the same time, there are differences in the interaction methods in different scenarios, such as the differences in contact patterns caused by holding devices with one or both hands, which may cause disturbances in the spatiotemporal distribution of HW-VIP.

To comprehensively evaluate the stability of *VIP* in various usage scenarios, we designed the above three typical usage scenarios. We adopted two common grip methods, including 1) gesture 1: holding with both hands, and 2) gesture 2: holding with one hand. As shown in Figure 12, *VIP* has the best authentication performance in static laboratory scenes (i.e., sitting), with an average F1-score of 98.2%. However, in dynamic scenes, especially when walking on stairs, it

TABLE II
COMPARISON OF AUTHENTICATION PERFORMANCE UNDER DIFFERENT OCCLUSION CONDITIONS.

	Accuracy	Precision	F1-score
With Occlusion	96.03%	95.73%	96.80%
Without Occlusion	98.47%	98.58%	98.22%

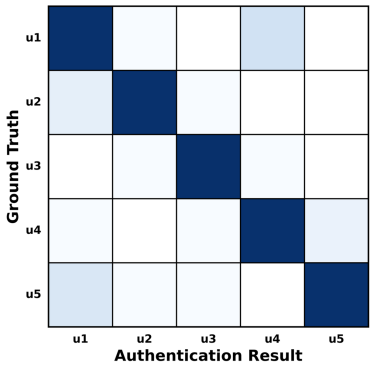


Fig. 14. Confusion matrices for the five participants.

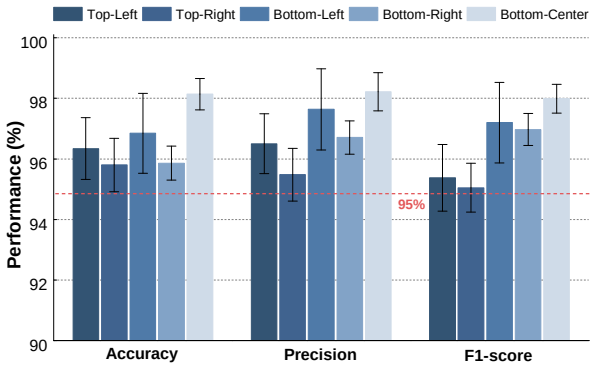


Fig. 15. Performance under different touch positions.

can cause strong hand tremors and irregular swings. The severe changes in hand acceleration and unstructured posture disturbances can interfere with signal acquisition, resulting in a decrease in F1-score to 94.5%. In addition, the authentication performance of dual-handed grip is better than that of single-handed grip, indicating that dual-handed grip is more stable and beneficial for improving the distinguishability of HW-VIP.

It is worth noting that although there are differences in the conditions of each scenario, *VIP* maintains F1-score above 94% in all combination settings. These results highlight *VIP*'s stable response to environmental and behavioral variations, underscoring its practicality in mobile identity authentication.

E. Impact of Hand Conditions (Q3)

To evaluate the identity authentication performance of *VIP* under different hand physiological states, we designed experiments covering four common hand conditions, including clean, slightly sweaty, wet, and dusty. As shown in Figure 13, the system performance significantly decreases in both "wet" and "dusty" states, with the FAR of *VIP* increasing to 2.55% and 1.35%, respectively. Compared to 1.48% in the clean state, the system performance has decreased to a certain extent in the wet hand state, and the FRR of *VIP* has increased to 6.13%.

This phenomenon is mainly due to HW-VIP relying on mechanical contact between fingers and screens to excite vibration signals. When the hands are in a non-clean state, the contact conditions between the skin and the device surface will undergo irregular changes. This phenomenon is manifested as a decrease in the elastic modulus of the contact interface and a drift in the damping coefficient, which weakens the energy coupling efficiency during vibration excitation (see §II-A) and causes distortion of the signals received by the IMU.

Especially when the fingers are wet, the dielectric layer formed by the water film further weakens the coupling and conduction of vibration signals, disrupting the repeatability and structural consistency of HW-VIP. To further expand the applicability boundaries of the system in various hand states, it may be considered to construct a dataset with state annotations in the future. We will explore conditional feature modeling methods based on state perception mechanisms to enhance the interpretability and generalization depth of the system.

F. Impact of Occlusion (Q3)

In daily use, the user's sleeve may partially cover the watch, causing unexpected noise and device displacement. To simulate the impact of clothing-induced occlusion on *VIP* performance in daily use, we randomly selected 5 volunteers from 30 legitimate users for the experiment. Participants are required to complete standardized interactive tasks with sleeves covering the watch, which are the same as the unobstructed conditions, to avoid interference caused by operational differences. As shown in Table II, compared with the unobstructed state, the average accuracy, precision, and F1-score of the system in this situation decreased by 2.44%, 2.85%, and 1.42%, respectively. The performance degradation is caused by occlusion from clothing interfering with the propagation path of the vibration response in the forearm area, resulting in local absorption or propagation distortion of the signal, which in turn affects the model's perception ability of key structural areas.

However, as shown in the confusion matrix in Figure 14, the system can still maintain a low error rate under occlusion conditions. This is attributed to D-BAFB's flexible correction of key point position information on the table side through learnable offset vectors, effectively avoiding local failures caused by local occlusion while still retaining key identity information. To improve *VIP*'s adaptability to complex occlusions, such samples can be incorporated during training to enhance robustness against disturbances like cuffs.

G. Impact of User Operation Behavior (Q3)

To systematically evaluate the performance of *VIP* under different operational behaviors, we further study the impact of common touch behaviors (e.g., changes in touch position and touch intensity) of users on authentication effectiveness during actual use. Referring to the typical grip habits of users, we have set up five representative interaction areas, which are

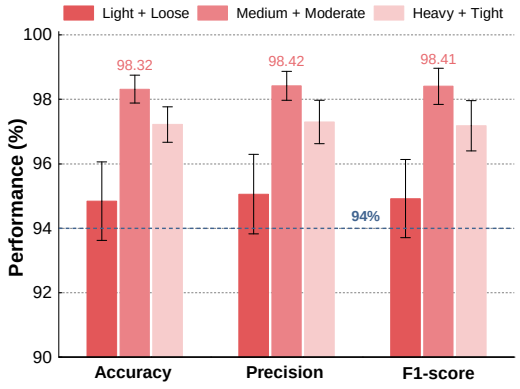


Fig. 16. Impact of touch force and wear tightness.

located in the top left, top right, bottom left, bottom right, and bottom center areas of the screen as shown in Figure 8.

As shown in Figure 15, although users touch at different screen positions, the system maintains over 95% accuracy, precision, and F1-score. Further observation reveals that the authentication effect in the lower area of the screen is better than that in the top area. Especially at the bottom-center position, the accuracy is 98.14%. The reason may be that the vibration motor is usually built into the middle of the bottom of the phone (see Figure 8). If the finger is close to the motor position, the vibration signal can be transmitted more efficiently to the hand and wrist, thereby enhancing the strength and consistency of HW-VIP. In contrast, the signal strength attenuation in the top region is more significant due to its distance from the vibration source, resulting in a decrease in the discriminative performance of this region. Overall, the performance differences between regions are relatively small, and the system has good stability in spatial distribution.

In terms of contact intensity, we constructed three typical interaction modes, including "light touch + loose fitting", "medium touch + moderate fitting", and "heavy touch + tight fitting", to simulate the interference of different usage states on sensor response. As shown in Figure 16, under three different combination conditions, *VIP* exhibits good authentication performance, with accuracy, precision, and F1-score remaining above 94%. Further analysis demonstrates that the combination of medium touch and moderate fitting can achieve optimal performance, with an average of over 98% for all three indicators.

In addition, unstable contact between the watch and the skin may cause loose or slight wear, resulting in micro-motion noise and poor contact, which affects the stability of feature extraction. Although heavy touch and tight fitting usually help maintain stable authentication, excessive mechanical contraction may cause signal saturation or attenuation, finally affecting the interpretability of touch-vibration signals. The Table III indicates that all users are accustomed to using Medium force touch in their daily operations. This type of operation method can ensure stable screen triggering without causing discomfort or misoperation. In contrast, Light touch is less common among users, with some expressing concerns about not triggering successfully and needing to repeat the touch, which affects the smooth experience. Heavy touch, on the other hand, often occurs in brief high-intensity operation scenarios (e.g., clicking the confirm button), and is not a conventional interaction method.

TABLE III
STATISTICS ON TOUCH AND WEARING HABITS OF DIFFERENT USERS IN DAILY DEVICE USE. HERE, ○, ◑, AND ● REPRESENT ALMOST NEVER USED, OCCASIONALLY USED, AND FREQUENTLY USED, RESPECTIVELY.

User ID	Screen touch intensity			Watch wearing elasticity		
	Light	Medium	Heavy	Loose	Moderate	Tight
U1	○	●	◑	○	●	○
U2	○	●	○	○	●	○
U3	○	●	○	○	●	○
U4	○	●	◑	○	●	○
U5	○	●	◑	○	●	○
U6	○	●	○	○	●	○
U7	○	●	◑	○	●	◑
U8	○	●	○	○	●	◑
U9	○	●	◑	○	●	○
U10	○	●	○	○	●	◑
U11	○	●	◑	○	●	○
U12	○	●	◑	○	●	○
U13	○	●	○	○	●	◑
U14	○	●	◑	○	●	○
U15	○	●	○	○	●	○
U16	◑	●	◑	○	●	○
U17	○	●	○	○	●	◑
U18	○	●	◑	○	●	○
U19	○	●	◑	○	●	◑
U20	○	●	○	○	●	○
U21	○	●	◑	○	●	○
U22	○	●	○	○	●	○
U23	○	●	○	○	●	◑
U24	○	●	○	○	●	○
U25	○	●	◑	○	●	○
U26	○	●	○	○	●	○
U27	○	●	◑	○	●	○
U28	○	●	◑	○	●	○
U29	◑	●	○	◑	●	○
U30	◑	●	○	○	●	◑

In terms of the elasticity of the watch, users generally prefer to wear it in a moderate manner. In the survey, multiple users pointed out that wearing the device too loosely can easily cause shaking and dis placement, affecting the comfort of long-term wearing. Although wearing it tight helps the device to adhere closely to the skin, it is often perceived by users as having a strong sense of pressure and fatigue, and is often used in special scenarios such as sports, with limited frequency of use in real authentication scenarios.

Excitingly, the combination of Medium touch intensity and Moderate wearing not only best fits the user's natural usage habits, but also achieves optimal system performance (see Figure 16). This result confirms that *VIP* can achieve excellent authentication performance and have good natural deployment without deliberately changing user usage patterns.

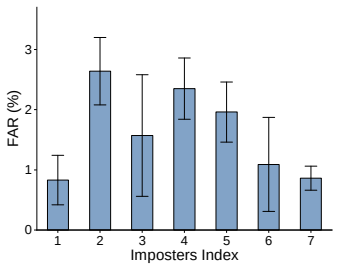


Fig. 17. FAR under external impersonation attacks.

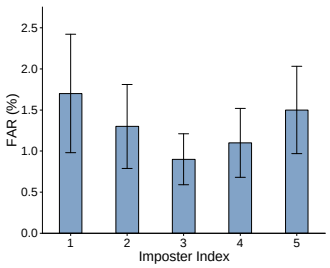


Fig. 18. FAR under internal impersonation attacks.

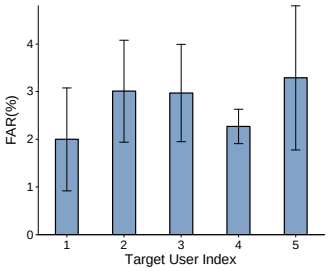


Fig. 19. FAR under side-channel replay attacks.

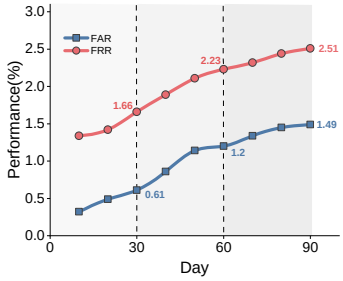


Fig. 20. Long-term authentication performance of *VIP*.

TABLE IV
COMPARISON OF *TouchFormer* PERFORMANCE WITH OR WITHOUT SAM IN DIFFERENT ATTENTION.

Method	With SAM			Without SAM		
	Accuracy	Precision	F1-score	Accuracy	Precision	F1-score
M-HA [35]	92.61%	92.90%	91.86%	91.25%	91.80%	90.65%
DDA [34]	96.74%	96.80%	96.88%	95.21%	95.46%	94.92%
DA [33]	96.29%	96.44%	96.40%	95.71%	95.86%	96.45%
D-BAFB	**98.65%**	**98.58%**	**98.22%**	**97.82%**	**98.17%**	**98.12%**

H. Ablation Study (Q4)

To evaluate the specific contribution of the introduction of D-BAFB and SAM optimization strategies to identity authentication performance, we compare the performance of different attention mechanisms with or without the introduction of SAM, including standard M-HA, DDA, DA, and the proposed D-BAFB. The experimental results are shown in Table IV. Among all settings, D-BAFB performs the best. After introducing SAM, its accuracy, precision, and F1-score reached 98.65%, 98.58%, and 98.22%, respectively, significantly better than other structures, reflecting its stronger feature modeling and discrimination ability. It is worth noting that even without introducing SAM, D-BAFB still maintains a leading position with an accuracy of 97.82% and an F1-score of 98.12%, indicating that its structure itself has good generalization.

Further analysis of the performance differences among various attention mechanisms reveals that: 1) The performance of M-HA is weak, especially in the absence of SAM, where F1-score is only 90.65%, indicating that this mechanism has limitations in handling key temporal differences in tactile sequences; 2) DDA can effectively model the temporal misalignment between mobile phones and watches through the dynamic attention mechanism controlled by λ_{dyn}. However, due to the lack of spatial alignment mechanism, the response is unstable under spatial disturbances; 3) DA can adaptively sample key touch positions through offset sensing mechanism. However, its modeling of response position differences caused by small disturbances still has certain shortcomings due to the lack of direction discrimination mechanism; and 4) D-BAFB achieves complementary advantages in modeling dynamic response and spatially sparse features. The fusion of the two attention outputs (see Figure 6) improves feature selectivity and adaptability to noise.

In addition, the SAM optimizer [41] brings F1-score improvement of 1% to 2% in each structure, enhancing the stability of the model under state disturbances and individual differences. This is due to its two-stage perturbation mechanism. By guiding the model to perform adversarial perturbations in the direction of greater loss curvature, it suppresses convergence to sharp minima while promoting parameter updates towards flatter minima. Intuitively, SAM performs better when *TouchFormer* exhibits higher loss sharpness.

I. Performance on Attack Resistance (Q5)

To evaluate the anti-attack performance of *VIP*, we conducted a series of experiments under two attack models:

1) Resist Impersonation Attacks: We randomly selected 5 volunteers (including 2 females and 3 males) from legitimate users as target users, and recruited 7 unregistered users as attackers to participate in the experiment (see §V-A). Specifically, each legitimate user first completes the registration process and selects the grip posture and touch method according to their own habits. Subsequently, the attacker is allowed to select legitimate users with similar physiological characteristics to themselves, and then repeatedly observes the authentication process of legitimate users to improve the success rate of the attack. Each imposter is allowed to practice before performing 20 authentication attempts.

For external impersonation attacks, attackers do not register in the system and only imitate by observing the touch behavior of legitimate users (see §III). As shown in Figure 17, in all experiments where attackers attempt to impersonate legitimate users, the average FAR for each attacker remains at 1.59%.

For internal impersonation attacks, we simulate unauthorized login when multiple users share devices. Every legitimate user has registered their identity in the system, but attackers attempt to impersonate other registered users in the system and gain the privileges that other users possess. Figure 18 demonstrates that when five legitimate users attempt to log in to other accounts as attackers, the average FAR is 1.50%.

2) Resist Replay Attacks: When legitimate users use smartphones, we allow imposters to use high-sensitivity external IMU devices (see Figure 4(a)) to collect vibration signals generated during the interaction between legitimate users and devices in the same desktop environment, and save them as attack samples. Subsequently, the attacker replayed the captured signal during the login phase, attempting to deceive *VIP* to complete identity verification. To comprehensively

evaluate the effectiveness of the attack, we allowed the attacker to repeatedly perform replay attacks on five legitimate users. As shown in Figure 19, despite the attacker's mastery of high-fidelity capture and playback techniques, *VIP* successfully identified and prevented most replay attacks with a lower FAR (i.e., an average of 2.71%).

J. Long-term Performance (Q6)

To evaluate the long-term availability of the *VIP* system, we conducted a 90-day experiment. As shown in Figure 20, even if the user's physiological state and behavioral habits undergo natural changes, the FAR and FRR of the system still remain at 1.49% and 2.51%, respectively. The overall growth rate is controlled within 1.2%, reflecting *VIP*'s good ability to resist temporal degradation. To further enhance long-term applicability, the system can combine individual temporal modeling with incremental update mechanisms. Continuous optimization of the model can address the natural drift of feature distribution.

In addition, we synchronously recorded the average authentication time required for each login to evaluate the practicality performance of the system. In 95% of authentication operations, *VIP* completes identity authentication within an average of 238 ± 41 ms, meeting users' requirements for speed and experience in daily scenarios.

VI. Related Work

A. Biometric-based Authentication

Biometric authentication usually relies on the user's inherent physical characteristics, such as fingerprints [7], irises [15], faces [4], and voiceprints [42], but these methods are vulnerable to replay attacks [10]. In addition to common methods, Pandia *et al*. [43] used a mobile phone camera to capture images of the user's teeth for authentication, which is greatly affected by lighting and shooting angles. Then, Xie *et al*. [44] used the built-in microphone of earplugs to collect bone conduction sound signals generated when teeth bite for verification. After that, Zhou *et al*. [45] utilized the friction sound of fingertips during the sliding unlocking process as the second factor for authentication. Srivastava *et al*. [46] identified the user by detecting the movement pattern of the jaw bone and facial micro-vibrations when the user speaks, but requires wearing special headphones to obtain IMU data.

B. Behavior-based Authentication

This type of method implements identity authentication based on the dynamic behavior characteristics of the user's interaction with the device, such as typing touch behavior [13] and walking gait [11]. For example, Yang *et al*. [47] designed a lightweight authentication system CALL, which is modeled based on the time series collected by the built-in sensors of the mobile phone. Li *et al*. [48] used the accelerometer and gyroscope in the mobile phone to collect the action sequence of the user's daily operation, and combined the autoencoder and discriminator to build the AEGAN model to complete the authentication. Wu *et al.* [49] achieved authentication by analyzing the acceleration and rotation angle during the touch process of the mobile phone.

C. Vibration-based Authentication

Identity authentication methods based on vibration signals have gradually become a research hotspot. For example, Li *et al*. [50] embeded a pair of linear vibration motors and IMU in VR/AR devices, and identified the vibration response pattern based on the head structure. Liu *et al*. [51] built a desktop authentication system that verifies through vibration sensing at a specific location, but relies on dedicated hardware. Yang *et al*. [22] and Lee *et al*. [52] found single-ended authentication schemes based on the built-in motor and IMU of a smartwatch, which uses the differential signal of vibration propagation in the wrist for identity authentication. Cao *et al*. [53] authenticated through the passive response of the device to vibration when the user holds the phone. Xu *et al*. [21] captured the vibration characteristics of the user's finger when touching the screen, while Xie *et al*. [23] focused on sliding-based interaction for identity authentication. The existing solutions focus on single-terminal authentication for smartphones or smartwatches, ignoring the cross-part transmission and coupling characteristics of vibration signals between the hand and wrist. Differently, *VIP* utilizes vibration motors and IMU sensors commonly found in most devices to construct a dual-terminal collaborative active vibration mechanism, utilizing the user's hand-wrist collaborative vibration mode under active vibration stimulation to achieve identity authentication.

VII. Discussion

Although *VIP* exhibits promising authentication performance across multiple devices, it still faces challenges in deployment and long-term use.

First, although we have validated the effectiveness of *VIP* in a user population covering multiple age groups and physiological structures, further investigation is needed on its adaptability and generalization ability in a wider range of populations, such as different races, body types, and body fat ratios. Especially, the feature boundary clarity of HW-VIP in complex population distributions are worthy of further research.

Second, the design of *VIP* is based on the built-in motors of smart devices. Although all participants reported that the vibration is non-disruptive, future work could further reduce user-perceived workload and enhance system concealment. Potential approaches include shortening the duration of stimulation and exploring stimulation strategies with lower salience.

Third, considering long-term usage scenarios, an individual's physiological state (e.g., bone density and muscle tone) and behavioral characteristics may slowly evolve over time, leading to drift in the HW-VIP distribution. In the future, a personalized model update strategy that is time-sensitive can be introduced to maintain stable model performance as the user's state dynamically evolves.

VIII. Conclusion

In this paper, we proposed a dual-terminal collaborative identity authentication system, named *VIP*. We deployed and evaluated the performance of *VIP* on various smart terminals. The experimental results demonstrated that our proposed *VIP*

outperforms existing competing authentication methods. Excitingly, we have found that: 1) *VIP* achieves optimal performance under touch intensity and suitable wearing conditions that conform to users' daily usage habits; and 2) *VIP* has demonstrated stable and consistent performance on devices of different brands in various authentication scenarios. In simulated and replay attack scenarios, additionally, the false acceptance rate of *VIP* is lower than that of the comparison methods, which verifies its adaptability and security in actual environments. In our future work, we will focus on more complex identity authentication scenarios. We will further explore the joint certification of various smart terminal devices such as smart headphones and smart bracelets.

REFERENCES

[1] S. Gill, "How many people own smartphones in the world?" https://prioridata.com/data/smartphone-stats/, 2025, [Online; accessed 07-August-2025].

[2] L. Hulsey, "2023 will go down for record-setting number of data breaches," https://www.governing.com/management-and-administration/2023-will-go-down-for-record-setting-number-of-data-breaches, 2024, [Online; accessed 07-August-2025].

[3] L. Lamport, "Password authentication with insecure communication," *Communications of the ACM*, vol. 24, no. 11, pp. 770–772, 1981.

[4] M. Guerar, M. Migliardi, F. Palmieri, L. Verderame, and A. Merlo, "Securing pin-based authentication in smartwatches with just two gestures," *Concurrency and Computation: Practice and Experience*, vol. 32, no. 18, p. e5549, 2020.

[5] N. Kausar, I. U. Din, M. A. Khan, A. Almogren, and B.-S. Kim, "Grapin: A graphical and pin-based hybrid authentication approach for smart devices," *Sensors*, vol. 22, no. 4, p. 1349, 2022.

[6] B. Zhou, Z. Xie, Y. Zhang, J. Lohokare, R. Gao, and F. Ye, "Robust human face authentication leveraging acoustic sensing on smartphones," *IEEE Transactions on Mobile Computing*, vol. 21, no. 8, pp. 3009–3023, 2021.

[7] A. S. Rathore, W. Zhu, A. Daiyan, C. Xu, K. Wang, F. Lin, K. Ren, and W. Xu, "Sonicprint: A generally adoptable and secure fingerprint biometrics in smart devices," in *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*, 2020, pp. 121–134.

[8] B. J. Tang and K. G. Shin, "Eye-shield: Real-time protection of mobile device screen information from shoulder surfing," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 5449–5466.

[9] Z. Muhammad, Z. Anwar, A. R. Javed, B. Saleem, S. Abbas, and T. R. Gadekallu, "Smartphone security and privacy: a survey on apts, sensor-based attacks, side-channel attacks, google play attacks, and defenses," *Technologies*, vol. 11, no. 3, p. 76, 2023.

[10] R. Gupta and P. Sehgal, "A complete end-to-end system for iris recognition to mitigate replay and template attack," in *Soft Computing and Signal Processing*, 2019, pp. 571–582.

[11] C. Wan, L. Wang, and V. V. Phoha, "A survey on gait recognition," *ACM Computing Surveys (CSUR)*, vol. 51, no. 5, pp. 1–35, 2018.

[12] J. H. Huh, S. Kwag, I. Kim, A. Popov, Y. Park, G. Cho, J. Lee, H. Kim, and C.-H. Lee, "On the long-term effects of continuous keystroke authentication: Keeping user frustration low through behavior adaptation," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 2, pp. 1–32, 2023.

[13] P. Negi, P. Sharma, V. Jain, and B. Bahmani, "K-means++ vs. behavioral biometrics: One loop to rule them all." in *NDSS*, 2018, pp. 1–13.

[14] A. Verma, V. Moghaddam, and A. Anwar, "Data-driven behavioural biometrics for continuous and adaptive user verification using smartphone and smartwatch," *Sustainability*, vol. 14, no. 12, p. 7362, 2022.

[15] W. Li, J. Wang, G. Zhang, Y. Yang, R. Spolaor, X. Cheng, and P. Hu, "Emiris: Eavesdropping on iris information via electromagnetic side channel." in *NDSS*, 2025, pp. 1–13.

[16] F. Lin, C. Song, Y. Zhuang, W. Xu, C. Li, and K. Ren, "Cardiac scan: A non-contact and continuous heart-based user authentication system," in *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, 2017, pp. 315–328.

[17] C. Wu, J. Chen, K. He, Z. Zhao, R. Du, and C. Zhang, "Echohand: High accuracy and presentation attack resistant hand authentication on commodity mobile devices," in *Proceedings of the 2022 ACM SIGSAC conference on computer and communications security*, 2022, pp. 2931–2945.

[18] L. Yang, W. Wang, and Q. Zhang, "Secret from muscle: Enabling secure pairing with electromyography," in *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*, 2016, pp. 28–41.

[19] C. Zippenfennig, B. Wynands, and T. L. Milani, "Vibration perception thresholds of skin mechanoreceptors are influenced by different contact forces," *Journal of Clinical Medicine*, vol. 10, no. 14, p. 3083, 2021.

[20] C. E. Connor, S. S. Hsiao, J. R. Phillips, and K. O. Johnson, "Tactile roughness: neural codes that account for psychophysical magnitude estimates," *Journal of Neuroscience*, vol. 10, no. 12, pp. 3823–3836, 1990.

[21] X. Xu, J. Yu, Y. Chen, Q. Hua, Y. Zhu, Y.-C. Chen, and M. Li, "Touchpass: Towards behavior-irrelevant on-touch user authentication on smartphones leveraging vibrations," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020, pp. 1–13.

[22] L. Yang, W. Wang, and Q. Zhang, "Vibid: User identification through bio-vibrometry," in *2016 15th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, 2016, pp. 1–12.

[23] Y. Xie, F. Li, and Y. Wang, "Fingerslid: Towards finger-sliding continuous authentication on smart devices via vibration," *IEEE Transactions on Mobile Computing*, vol. 23, no. 5, pp. 6045–6059, 2023.

[24] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," *International Conference on Learning Representations*, pp. 1–19, 2021.

[25] A. F. Cristea and R. Morariu-Gligor, "The simulation of vibration attenuation in the hand-arm sistem," *UPB Scientific Bulletin, Series D*, vol. 75, no. 3, pp. 1–15, 2013.

[26] E. Shahabpoor, A. Pavic, and V. Racic, "Identification of mass–spring–damper model of walking humans," in *Structures*, vol. 5, no. 1, 2016, pp. 233–246.

[27] W. Klippel, "Tutorial: Loudspeaker nonlinearities—causes, parameters, symptoms," *Journal of the Audio Engineering Society*, vol. 54, no. 10, pp. 907–939, 2006.

[28] P. Duhamel and M. Vetterli, "Fast fourier transforms: a tutorial review and a state of the art," *Signal processing*, vol. 19, no. 4, pp. 259–299, 1990.

[29] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. 1, pp. 2579–2605, 2008.

[30] C. Wang, Y. Wang, Y. Chen, H. Liu, and J. Liu, "User authentication on mobile devices: Approaches, threats and trends," *Computer Networks*, vol. 170, no. 1, p. 107118, 2020.

[31] M. Singh and D. Pati, "Countermeasures to replay attacks: A review," *IETE Technical Review*, vol. 37, no. 6, pp. 599–614, 2020.

[32] G. Zhao, Q. Jiang, X. Liu, X. Ma, N. Zhang, and J. Ma, "Electrocardiogram based group device pairing for wearables," *IEEE Transactions on Mobile Computing*, vol. 22, no. 11, pp. 6394–6409, 2022.

[33] Z. Zhou, G. Lyu, Y. Huang, Z. Wang, Z. Jia, and Z. Yang, "Sdformer: transformer with spectral filter and dynamic attention for multivariate time series long-term forecasting," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24), Jeju, Republic of Korea*, 2024, pp. 3–9.

[34] D. Luo and X. Wang, "Deformabletst: Transformer for time series forecasting without over-reliance on patching," in *Neural Information Processing Systems*, 2024, pp. 88 003–88 044.

[35] 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, no. 1, pp. 1–15, 2017.

[36] A. Y. Kei and S. S. Chow, "Shaft: Secure, handy, accurate, and fast transformer inference," in *Network and Distributed System Security Symposium, NDSS*, 2025, pp. 1–14.

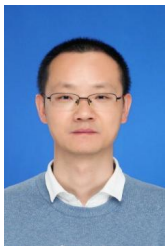
[37] G. Lai, W.-C. Chang, Y. Yang, and H. Liu, "Modeling long-and short-term temporal patterns with deep neural networks," in *The 41st international ACM SIGIR conference on research & development in information retrieval*, 2018, pp. 95–104.

[38] A. Dempster, F. Petitjean, and G. I. Webb, "Rocket: exceptionally fast and accurate time series classification using random convolutional kernels," *Data Mining and Knowledge Discovery*, vol. 34, no. 5, pp. 1454–1495, 2020.

[39] D. Campos, M. Zhang, B. Yang, T. Kieu, C. Guo, and C. S. Jensen, "Lightts: Lightweight time series classification with adaptive ensemble

distillation," *Proceedings of the ACM on Management of Data*, vol. 1, no. 2, pp. 1–27, 2023.
[40] Y. Zhang and J. Yan, "Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting," in *The Eleventh International Conference on Learning Representations*, 2023, pp. 1–21.
[41] R. Ilbert, A. Odonnat, V. Feofanov, A. Virmaux, G. Paolo, T. Palpanas, and I. Redko, "Samformer: Unlocking the potential of transformers in time series forecasting with sharpness-aware minimization and channel-wise attention," in *Forty-first International Conference on Machine Learning*, 2024, pp. 1–31.
[42] R. Zhang, Z. Yan, X. Wang, and R. H. Deng, "Livoauth: Liveness detection in voiceprint authentication with random challenges and detection modes," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 6, pp. 7676–7688, 2022.
[43] A. Pandia, G. Arora, A. Jain, R. Bharadwaj, A. Bhatia, and K. Tiwari, "Dteeth: Teeth-photo based human authentication for mobile devices," in *2022 IEEE International Joint Conference on Biometrics (IJCB)*, 2022, pp. 1–8.
[44] Y. Xie, F. Li, Y. Wu, H. Chen, Z. Zhao, and Y. Wang, "Teethpass: Dental occlusion-based user authentication via in-ear acoustic sensing," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, 2022, pp. 1789–1798.
[45] M. Zhou, Y. Zhou, S. Su, Q. Wang, Q. Li, S. Hu, C. Yu, and Z. Li, "Fingerpattern: Securing pattern lock via fingerprint-dependent friction sound," *IEEE Transactions on Mobile Computing*, vol. 23, no. 6, pp. 7210–7224, 2023.
[46] T. Srivastava, S. Pan, P. Nguyen, and S. Jain, "Jawthenticate: Microphone-free speech-based authentication using jaw motion and facial vibrations," in *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*, 2023, pp. 209–222.
[47] Z. Yang, Y. Li, and G. Zhou, "Unsupervised sensor-based continuous authentication with low-rank transformer using learning-to-rank algorithms," *IEEE Transactions on Mobile Computing*, vol. 23, no. 9, pp. 8839–8854, 2024.
[48] Y. Li, C. Ouyang, and H. Huang, "Aeganauth: Autoencoder gan-based continuous authentication with conditional variational autoencoder generative adversarial network," *IEEE Internet of Things Journal*, 2024.
[49] C. Wu, K. He, J. Chen, Z. Zhao, and R. Du, "Liveness is not enough: Enhancing fingerprint authentication with behavioral biometrics to defeat puppet attacks," in *29th USENIX Security Symposium (USENIX Security 20)*, 2020, pp. 2219–2236.
[50] F. Li, J. Zhao, H. Yang, D. Yu, Y. Zhou, and Y. Shen, "Vibhead: An authentication scheme for smart headsets through vibration," *ACM Transactions on Sensor Networks*, vol. 20, no. 4, pp. 1–21, 2024.
[51] J. Liu, C. Wang, Y. Chen, and N. Saxena, "Vibwrite: Towards finger-input authentication on ubiquitous surfaces via physical vibration," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 73–87.
[52] S. Lee, W. Choi, and D. H. Lee, "The vibration knows who you are! a further analysis on usable authentication for smartwatch users," *Computers & Security*, vol. 125, no. 1, p. 103040, 2023.
[53] H. Cao, H. Jiang, K. Yang, S. Chen, W. Wu, J. Liu, and S. Dustdar, "Data-augmentation-enabled continuous user authentication via passive vibration response," *IEEE Internet of Things Journal*, vol. 10, no. 16, pp. 14 137–14 151, 2023.

Zicheng Cui is currently pursuing a Master's degree in Electronic Information at Chang'an University, Xi'an, China. His current research interests include mobile computing, human-computer interaction, wearable devices, and acoustic side-channel security based on keyboard sounds.



Zhihai Yang received the Ph.D. degree in Control Science and Engineering from Xi'an Jiaotong University, China, in 2016. He is currently a professor with the School of Data Science and Artificial Intelligence, Chang'an University, Xi'an, China. His research interests include artificial intelligence security, identity security authentication, and cognitive security assessment.

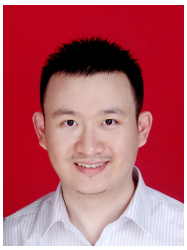
Zhiquan He is currently pursuing a Master's degree in Cyberspace Security at Chang'an University, Xi'an, China. His current research interests include mobile computing, human-computer interaction, wearable devices, and acoustic side-channel security based on keyboard sounds.

Chenxu Kong is currently pursuing a Master's degree in Cyberspace Security at Chang'an University, Xi'an, China. His current research interests include Backdoor Attacks on Computer Vision Models and Artificial Intelligence Security.

Jianhua He received the Ph.D. degree from Nanyang Technological University. He is currently a Professor with the University of Essex, U.K. He published more than 200 research articles. His research interests include data analytics and recommendation systems, computer security, mobile networking and computing, 5G/6G networks, Internet of Things, edge computing and intelligence, connected autonomous driving, intelligent transport systems, AI and large language models. He served as an Editor for IEEE Wireless Communication Letters and The Computer Journal.

Jianxin Li (Senior Member, IEEE) received the PhD degree in computer science from the Swinburne University of Technology, Australia, in 2009. He is professor of information systems with the School of Business and Law, Edith Cowan University, Australia. His research interests include graph database query processing and optimization, social network analytics and computing, complex network representation learning, and personalized online learning analytics. He is also a grant assessor in Australia Research Council Discovery Programs and Linkage Programs, and serves as invited reviewers for many top journals and program committee members in many top conferences.

Pinghui Wang (Senior Member, IEEE) is currently a Professor with the MOE Key Laboratory for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an, China, and also with the Shenzhen Research Institute, Xi'an Jiaotong University, Shenzhen, China. His research interests include internet traffic measurement and modeling, traffic classification, abnormal detection, and online social network measurement.



Zhiquan Liu received the B.S. degree from the School of Science, Xidian University, Xi'an, China, in 2012, and the Ph.D. degree from the School of Computer Science and Technology, Xidian University, Xi'an, China, in 2017. He is currently a full professor, doctoral supervisor, and deputy dean with the College of Cyber Security, Jinan University, Guangzhou, China. His current research focuses on security, trust, privacy, and intelligence in vehicular networks and UAV networks. He currently serves as the area editor or associate editor of multiple SCI-index journals, such as IEEE TIFS, IEEE TDSC, IEEE TII, IEEE TVT, IEEE IOTJ, IEEE Network, Information Fusion, etc. His homepage is https://www.zqliu.com.