Deep Learning for Virtual Reality User Identification: A Benchmark

Davide Frizzo[1,*], Fabrizio Genilotti[1,*], David Petrovic[1,*], Arianna Stropeni[1,*],
Francesco Borsatti[1], Davide Dalle Pezze[1], Riccardo De Monte[1],
Gian Antonio Susto[1]

***Abstract*— Virtual Reality (VR) applications require robust user identification systems to ensure secure access to equipment and protect worker identities. Motion tracking data from VR headsets and controllers has emerged as a powerful behavioral biometric, with recent studies demonstrating identification accuracies exceeding 94% across a large user base. However, the application of modern deep learning architectures, particularly State Space Models (SSM), to VR scenarios remains largely unexplored. In this work, we benchmark user identification performance across the large-scale *Who is Alyx* VR dataset, gathering data from 71 users playing the popular Half-Life:Alyx game. We evaluate both established architectures (Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Convolutional Neural Network (CNN), Temporal Convolutional Network (TCN), Transformer) and the emerging SSMs on time series motion data. Our results provide the first comprehensive benchmark of state-of-the-art and novel architectures for VR user identification, establishing baseline performance metrics for future privacy-preserving authentication systems in manufacturing environments.**

I. Introduction

The rapid adoption of VR technology across diverse sectors has created new challenges for user authentication and identity protection. Modern VR systems continuously collect rich behavioral data from Head-Mounted Display (HMD) and hand controllers, generating high-dimensional time series that capture individual movement patterns. Recent research has demonstrated that this motion data serves as a powerful biometric identifier, with classification accuracies approaching those of traditional biometric modalities like fingerprints [1].

Secure user identification in VR environments is critical across multiple application domains. In healthcare, VR-based surgical training and telemedicine platforms require robust authentication to ensure that only qualified medical professionals access patient data and perform remote procedures. In education, virtual classrooms and examination systems need identity verification to maintain academic integrity and prevent credential fraud. In manufacturing and industrial settings, access to dangerous equipment, such as forklifts, industrial robots, and machinery, must be restricted to authorized and properly trained personnel, while VR-based training systems serve as certification platforms where identity verification ensures training completion authenticity [2].

Furthermore, previous work has predominantly employed traditional machine learning with hand-crafted statistical features or established deep learning architectures (LSTMs, CNNs) from the mid-2010s. The recent emergence of SSMs such as Structured State Spaces (S4), Diagonal State Space (S4D) and Simplified Structured State Space Layers (S5) [3], [4], [5], which demonstrate superior performance on long-sequence modeling tasks, has not been systematically evaluated for VR user identification.

This paper makes the following contributions:

- **Comprehensive benchmarking:** We evaluate user identification performance across the *Who is Alyx* dataset using both established deep learning architectures and modern SSMs.
- **Extension of feature-based methods:** End-to-end deep learning architectures with minimal pre-processing on the raw motion sequences are considered, in contrast with previous methods, mainly employing statistical feature engineering
- **State space model evaluation:** We provide the first systematic evaluation of structured state space models (S4D and S5) for VR user identification, comparing their performance and computational efficiency against traditional architectures.
- **Performance-efficiency analysis:** We provide an analysis of accuracy-efficiency trade-offs across architectures, establishing baseline metrics that enable informed model selection for VR authentication systems based on deployment constraints.

The remainder of this paper is organized as follows. Section II reviews related work in VR behavioral biometrics and time series classification. Section III describes our methodology, including the VR dataset characteristics (Section III-A) and the deep learning architectures evaluated in our benchmark (Section III-C), with emphasis on SSMs in Section III-B. Section IV describes our experimental setup. Section V presents results and discussion. Finally, Section VI concludes with future research directions.

II. Related Work

A. VR Behavioral Biometrics for User Identification.

Behavioral biometrics in VR has progressed from early proofs of concept to scalable identification systems. Pfeuffer et al. [6] demonstrated that body motion patterns and spatial relationships between tracked points (e.g., head–hand distance and inter-hand angles) are discriminative for user identification, showing that movements during basic VR tasks yield user-specific signatures. Identification research has scaled to large populations with Nair et al. [1], who

[1]University of Padua, Padua, Italy
[*]Authors contributed equally to this work

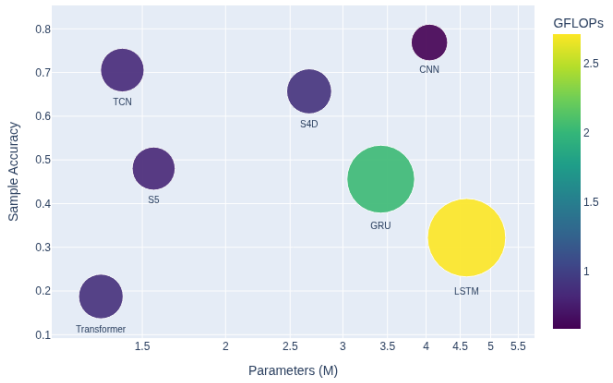


Fig. 1: Performance and efficiency trade-offs: comparison of deep learning models across parameter count (millions), sample accuracy, and computational complexity (GFLOPS).

analyzed motion data from 55,541 Beat Saber players and achieved up to 94.33% accuracy using hierarchical classification, with gradient boosting outperforming deep models, indicating that feature-engineered approaches can surpass end-to-end deep learning in this setting.

Despite these advances, behavioral biometrics face temporal stability challenges: Liebers et al. [7] showed that body normalization improves cross-session identification by 38%, but their 2023 field study [8] found performance drops from 86% to 71% over time, motivating dataset refresh cycles of roughly nine months. Cross-system generalization presents additional challenges. Miller et al. [9] achieved Equal Error Rates of 1.38–7.78% across different VR platforms (Oculus Quest, HTC Vive, Vive Cosmos) using Siamese neural networks. Their subsequent work [10] introduced scale and translation-invariant features that outperformed baselines in 36 of 42 experimental conditions. Most recently, Rack et al. [11] demonstrated 78.5% cross-application identification accuracy using transformer-based similarity learning across five different VR games.

Despite advances in VR behavioral biometrics, most existing studies rely on datasets where each user contributes only a few minutes of recording time. This data scarcity fundamentally limits the evaluation of deep learning architectures, which require substantial training samples to leverage their capacity for automatic feature learning from raw sequences. To properly benchmark deep learning architectures and assess their true potential for VR user identification, we selected *Who is Alyx*, a publicly available VR biometric dataset, containing a large per-user recording time (approximately 45 minutes per session across two sessions).

B. Time Series Classification Architectures.

Deep learning for time series classification has progressed through several architectural paradigms. The DeepConvLSTM architecture [12] established the CNN+LSTM hybrid approach, achieving state-of-the-art results on human activity recognition benchmarks by capturing both spatial features and temporal dynamics. InceptionTime [13] demonstrated that ensemble Inception networks achieve state-of-the-art performance on the UCR archive while being orders of magnitude faster than previous methods, training on 8 million time series in 13 hours.

Transformer architectures have been adapted for time series, with attention mechanisms enabling long-range dependency modeling. Recent work combines transformers with recurrent units, leveraging attention for global context and GRUs for local temporal patterns. However, transformers face quadratic complexity with respect to the sequence length. This poses a critical challenge in extended VR sessions.

SSM represent an emerging paradigm. The S4 architecture [14] solved the Path-X task (16,000 length sequences) where all prior models failed, with 60× faster generation than transformers. These results are attributable to the peculiar architecture of these models which enables parallel training similarly to CNN and unbounded context in inference as in Recurrent Neural Network (RNN).

III. Methodology

A. VR Scenario: Alyx Dataset

We evaluate user identification performance on the publicly available *Who is Alyx* VR dataset. This dataset [15] includes tracking data from 71 VR users (36 females and 40 males [1] between 21 and 33 years of age) playing Half-Life: Alyx on an HTC Vive Pro across two separate sessions. Each participant played for approximately 45 minutes per session, with sessions conducted on different days and on different chapters of the game to evaluate cross-session identification stability.

Half-Life: Alyx provides a rich contextualized environment where users perform diverse actions: navigation via teleportation or physical walking, object manipulation with gravity gloves, combat with weapon reloading, and puzzle solving with hand scanners and 3D puzzles. This diversity of interactions makes the dataset particularly challenging for identification, as input sequences contain arbitrary combinations of actions rather than repeated specific movements.

This dataset represents a significant advancement over prior VR behavioral biometric datasets in terms of scale. Each user contributed approximately 90 minutes of gameplay data across two sessions, providing substantially more training data per individual than previous collections. This extended per-user data volume is critical for deep learning approaches, as earlier datasets with shorter recording windows often lacked sufficient samples to effectively train complex neural network architectures.

Tracking Modalities: The dataset comprises three categories of tracking data: (1) motion data, including position

[1]The total number of users is 76 but only 71 of them participated in two sessions

and orientation of the head-mounted display and controllers; (2) gaze data from eye tracking sensors; and (3) physiological signals captured via an Empatica E4 wristband (acceleration, heart rate, electrodermal activity, photoplethysmography, inter-beat interval, and peripheral body temperature) and a Polar H10 chest strap (acceleration and electrocardiogram). The data are anonymized to guarantee users' privacy and satisfy ethical guidelines.

Key Challenges: This multimodal data presents several key challenges for user identification:

- High dimensionality: Modern VR systems track multiple points (e.g., HMD + 2 controllers = 21 features per frame), generating high-dimensional time series.
- Long sequences: VR sessions can span tens of minutes at 15-90 Hz sampling rates, producing sequences with tens of thousands of time steps.
- Cross-session generalization: Models must identify users from behaviors exhibited in different time periods than the training data, which is challenging due to the fact that user movements can vary across sessions due to task learning effects, fatigue, and environmental factors.

Temporal Encoding: To capture different aspects of user motion patterns, three temporal data encodings are evaluated, each emphasizing distinct kinematic properties and providing several levels of feature engineering:

- **Body Relative (BR) (Body-Relative):** Raw body-relative positions and orientations that preserve the absolute spatial configuration of controllers relative to the HMD reference frame.
- **Body Relative Velocity (BRV) (Body-Relative Velocity):** First-order temporal derivatives representing instantaneous motion dynamics. For positional data, frame-to-frame differences capture the velocity, while quaternion delta rotations encode angular velocity for orientations.
- **Body Relative Acceleration (BRA) (Body-Relative Acceleration):** Second-order temporal derivatives computed from BRV data, representing acceleration patterns. This encoding captures motion smoothness, jerk characteristics, and rapid directional changes that may reflect individual motor control strategies and reflexes during gameplay.

Cross-Domain Generalizability:

Beyond gaming applications, the Who is Alyx dataset serves as a valuable proxy for evaluating user identification systems across diverse VR domains such as healthcare and manufacturing. The variety of motor behaviors captured during gameplay, precision hand movements for object manipulation, coordinated hand-eye actions for targeting, and sustained attention during navigation has the potential to closely mirror the interaction patterns required in professional VR applications.

Consequently, identification models trained and benchmarked on this dataset provide meaningful insights into their potential performance in safety-critical environments where robust authentication is essential, such as restricting access to virtual surgical platforms or certifying authorized personnel for operating dangerous industrial equipment through VR-based training systems.

B. State Space Models

State Space Models represent the latest paradigm shift in sequence modeling architectures, emerging as a promising alternative to both recurrent and attention-based approaches. While established architectures like LSTMs, CNNs, and Transformers have dominated time series classification for the past decade, SSMs have only recently gained attention in deep learning as an efficient alternative to attention-based architectures for sequence modeling ([3]). Unlike Transformers, whose self-attention mechanism scales quadratically with sequence length, SSMs offer linear complexity, making them particularly well-suited for modeling long sequences such as the ones recorded in VR sessions.

Given their recent emergence and lack of evaluation in the VR biometrics domain, we dedicate specific attention to SSMs in this benchmark. Understanding whether these state-of-the-art sequence models can outperform established architectures on VR motion data have important implications for both identification accuracy and deployment feasibility on resource-constrained VR hardware.

An SSM models the evolution of an internal state $x(t)$ driven by an input signal $u(t)$ and producing an output $y(t)$:

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \\ y(t) = Cx(t) + Du(t) \end{cases} \tag{1}$$

where the system matrices A, B, C, and D are learned from data. In its discrete-time formulation, SSM closely resembles linear RNN, but lacks non-linear activation functions.

Such a linear structure equips SSM with a dual interpretation. On one hand, they can be viewed as recurrent models that update a hidden state over time, similar to RNN. On the other hand, unfolding the state dynamics reveals an equivalent convolutional form, where the output is obtained by convolving the input sequence with a fixed kernel. This convolutional view enables parallel training via Fast Fourier Transform (FFT), combining the long-range dependency modeling of RNN with the computational efficiency of CNN.

Several variants of the SSM architecture have been proposed over the years [16] across different domains such as Natural Language Processing (NLP) speech recognition, vision, and time series forecasting.

However, the most commonly used SSM for time series modeling are `S4`, `S4D`, and `S5` ([3],[4],[5]).

The `S4` model leverages the dual recurrent-convolutional nature of SSM and is designed as a Single Input Single Output (SISO) system, where each input feature is processed independently by a separate SSM block and combined through a mixing layer. `S4D` simplifies the architecture by restricting the system matrices to be diagonal, reducing computational cost at the expense of theoretical expressiveness. The S5 model further simplifies S4 by adopting a Multi Input Multi

Output (MIMO) formulation, allowing all input features to be processed jointly within a single SSM block.

C. Model Architectures

We conduct a comprehensive benchmark of deep learning architectures spanning several classes of sequence modeling approaches. Our evaluation includes traditional recurrent networks (LSTM, GRU), convolutional approaches (CNN, TCN), attention-based models (Transformer), and the recently introduced State Space Models (S4D, S5). This diverse architectural comparison enables us to identify which inductive biases, whether convolutional locality, recurrent temporal dynamics, global attention, or structured state space representations, are most effective for capturing discriminative motion patterns in VR behavioral biometrics. Furthermore, we are going to analyze the accuracy-efficiency trade-offs achieved by each model, discussing the feasibility of deployment for each model on resource-constrained VR hardware.

The evaluated architectures are:

- Multi-Layer Perceptron (MLP): A feedforward neural network trained on hand-crafted statistical features (min, max, mean, quantiles, and standard deviation). It is used as a baseline to assess the value of end-to-end deep learning approaches versus traditional feature engineering methods.
- Convolutional Neural Network (CNN): It applies convolutional filters to extract local temporal patterns from motion sequences, enabling the model to learn motion primitives at different temporal scales.
- Long Short-Term Memory (LSTM): A recurrent architecture with gating mechanisms designed to capture long-term dependencies.
- Gated Recurrent Unit (GRU): A simplified recurrent architecture that reduces the computational complexity while maintaining the ability to model temporal dependencies in motion sequences.
- Temporal Convolutional Network (TCN): Employs dilated causal convolutions to achieve large receptive fields while maintaining parameter efficiency. The dilated convolutions enable the model to capture long-range temporal patterns without the sequential computation constraints of recurrent models.
- Transformer: It utilizes multi-head self-attention mechanisms to model dependencies across the entire input sequence.
- Diagonal State Space (S4D): A structured state space model that restricts system matrices to diagonal form, reducing S4's computational cost while maintaining efficient long-sequence modeling.
- Simplified State Space Layer (S5): Further simplifies the SSM architecture by adopting a Multi-Input Multi-Output (MIMO) formulation, allowing all input features to be processed jointly within a single SSM block.

IV. Experimental Setup

A. Alyx Dataset

The original benchmark study [15] evaluated on CNN and GRU architectures. Our work extends this benchmark by evaluating other state-of-the-art approaches, including SSM, on the same data. Since the dataset has two separate sessions for each user, models are trained on session 1 and evaluated on session 2 for all users.

1) Data Preprocessing: The following preprocessing steps were considered:

- **Coordinate System Transformation:** Following Rack et al. [15], we transform raw scene-relative coordinates to body-relative coordinates using the Head-Mounted Display (HMD) as the reference frame. This removes positional and absolute orientation information that would allow models to overfit to spatial locations rather than movement patterns.
- **Downsampling to 15 fps**: Since the majority of the sessions are recorded at 15 fps we decided to keep this as the sampling frequency of the time series signals by downsampling the data from the 90 fps sessions.
- **Removal of noisy samples**: The first and last minute of each session are removed to discard noisy samples.

Each encoding results in 18 features per frame: 7 features per controller (position x,y,z; rotation quaternion x,y,z,w) plus 4 features for HMD rotation (pitch and yaw only; roll and position become constant in body-relative coordinates).

2) Sequence Windowing: We segment continuous recordings into fixed-length input sequences. In particular, we use 20-second windows (300 frames at 15 Hz), consistent with the original benchmark. Sequences are sampled with 50% overlap during training to increase the dataset size. During evaluation, we use non-overlapping windows and apply majority voting across all windows in a test recording to produce a single user prediction per session.

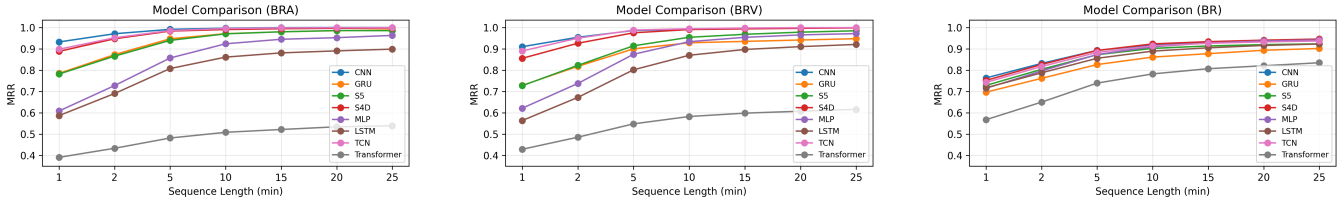
B. Training Configuration

Regarding the training configuration, the Adam optimizer is used with Early Stopping and a patience factor of 5 to mitigate overfitting. In particular, the set of model's parameters used for the inference evaluation is chosen by minimizing the mean accuracy on the validation set (i.e., the last 5 minutes of a training session).

C. Evaluation Metrics

Computation Complexity: Floating Point Operations (FLOP)s are considered to measure the computational complexity of a model as they provide a rough estimate of a model's training and inference time.

Performance: Taking into account the difficulty of the task, the MRR metric is considered at this stage. MRR quantifies the ranking quality of the model's predictions by measuring, for each sequence, the position at which the correct class appears in the ordered list of predicted classes. For each sequence window, the classes are ranked according to their



(a) BRA (b) BRV (c) BR

Fig. 2: Test sample size vs. Mean Reciprocal Rank (MRR) for the BRA, BRV and BR encodings

predicted scores, and the reciprocal of the rank of the ground-truth class is computed. The final MRR is obtained by averaging these reciprocal ranks over all sequences, providing a continuous measure of how consistently the correct class appears near the top of the ranked predictions.

Memory: Another considered metric in the study is the number of trainable parameters, since it is directly related to the memory required during the inference phase. Memory is a critical parameter when considering the deployment of DL models on standalone VR headsets with limited onboard memory.

Model	Parameters (M)	FLOPs (GFLOPs)
MLP	0.8935	0.0018
CNN	4.046	0.5886
GRU	3.42	2.0324
LSTM	4.6	2.7101
TCN	1.4	0.8394
Transformer	1.3	0.8793
S4D	2.67	0.8951
S5	1.56	0.818

TABLE I: Model parameters and Mult-Adds operations

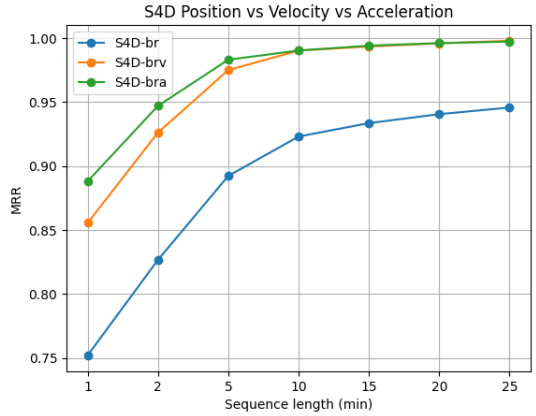


Fig. 3: Test sample size vs. MRR

V. Results

In this section, we highlight the experimental results obtained on the Who is Alyx dataset.

A. Performance Comparison Across Architectures

First, we evaluate the capabilities of deep learning models in VR User Identification.

In Figure 2 the MRR metric is evaluated in relation to the length of the time-series presented to the model at test time, showing the results based on the used encoding among BRA, BRV, and BR represented in 2a, 2b, and 2c respectively.

The first observation is that the MRR increases as the amount of test data increases for all data encodings, as majority voting across longer sequences reduces prediction noise and improves classification stability on the sequence. Moreover, comparing the curves associated with the different models, it is possible to conclude that CNN, TCN, and S4D are the most accurate ones, while LSTM and Transformer are at the bottom of the ranking. In particular, the Transformer model performs significantly worse than the other architectures.

This result is likely due to the limited dataset size and absence of large-scale pretraining. This suggests that the continuous-time nature of the VR sensor readings is more appropriate for models like TCN and SSM, rather than LSTM and Transformer models which are more suitable for discrete-time data typical of NLP tasks.

Finally, the Multi Layer Perceptron (MLP) model shows inferior results, placed between the best-performing models and the LSTM model. This result confirms the superiority of end-to-end deep learning approaches using sequential models rather than extracting statistical features from raw data. This analysis is confirmed in Figure 3, where the performances of the S4D model on the three different data encodings are compared.

B. Impact of Temporal Encoding Strategies

Among the three data encodings considered, the BRA encoding produced the best performances across all models in accordance with [15], followed by BRV and BR (as visible in Figure 3 for the S4D model). In particular, BRA and BRV show the most significant differences in MRR in the low test data regimes and saturate to overlapping results for sequences longer than 10 minutes. On the other hand, the BR encoding shows significantly worse results. In fact, velocity and acceleration information, obtained through consecutive sample differences on the raw position and velocity data, emphasize motion dynamics over absolute spatial positioning. These derivative-based features are invariant to constant offsets and linear drift in the tracking system, capturing how users move, which proves more discriminative for behavioral identification. Moreover, exploiting relative measurements

rather than absolute ones is a privacy-preserving feature which is crucial for VR-based authentication systems.

C. Memory and Computational Efficiency Considerations for Deployment

To conduct a final evaluation not only in terms of model performances but also with respect to efficiency and computational requirements, Figure 1 captures all three metrics in a single plot and shows the trade-offs across architectures. In particular, the horizontal axis contains the amount of model parameters (in millions), the model's sample accuracies in the BRA data encoding are reported in the vertical axis, and the size of the blob's radius is proportional to the model's FLOPs.

Examining the performance axis, convolutional-based architectures (CNN, TCN) and SSMs like S4D emerge as the top-performing models, significantly outperforming recurrent (LSTM, GRU) and attention-based (Transformer) approaches. In addition, the top three models (CNN, TCN, S4D) share similar computational requirements (FLOPs), indicating comparable training and inference times. However, CNN's high parameter count (4.0M) makes it memory-bound, potentially limiting deployment on standalone VR headsets with constrained RAM. TCN (1.4M parameters) and S4D (2.7M parameters) offer a better choice in this regard: they achieve only marginally lower accuracy than CNN while requiring 65% and 33% fewer parameters, respectively.

Therefore, TCN and S4D represent optimal choices when deployment constraints are considered, achieving competitive performance with substantially lower memory footprints.

VI. Conclusion

This study provided a comprehensive benchmark of deep learning architectures for user identification within VR contexts. By evaluating a spectrum of models ranging from traditional RNNs to modern SSMs on the *Who is Alyx* dataset, we established several key performance metrics and trade-offs.

The experimental results demonstrate that BRA encoding consistently provides the most discriminative features for identification across all architectures. Performance analysis reveals that CNN, TCN, and the S4D model represent the state-of-the-art for this task, significantly surpassing Transformers and recurrent models. Notably, while CNNs achieve high accuracy, their memory requirements may be prohibitive for certain applications. In contrast, TCN and S4D architectures offer a superior balance between identification accuracy, parameter efficiency, and computational demand, making them highly suitable for deployment on standalone VR hardware.

The robust user identification capabilities demonstrated in this benchmark have significant practical implications across multiple domains. In healthcare settings, reliable VR-based authentication can ensure that only qualified medical professionals access surgical training platforms and telemedicine systems. In industrial environments, accurate user identification enables secure access control to dangerous equipment such as industrial robots and forklifts. The performance-efficiency trade-offs established in this work can inform the deployment of authentication systems on resource-constrained VR hardware in real-world scenarios.

Future research directions include addressing cross-application generalization challenges for SSMs, studying the impact of user learning in noisy environments and adaptation on long-term identification accuracy, and developing efficient model compression techniques for edge deployment.

References

[1] V. Nair, W. Guo, J. Mattern, R. Wang, J. F. O'Brien, L. Rosenberg, and D. Song, "Unique identification of 50,000+ virtual reality users from head & hand motion data," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 895–910.

[2] G. M. Garrido, V. Nair, and D. Song, "Sok: Data privacy in virtual reality," *arXiv preprint arXiv:2301.05940*, 2023.

[3] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396*, 2021.

[4] A. Gupta, A. Gu, and J. Berant, "Diagonal state spaces are as effective as structured state spaces," 2022. [Online]. Available: https://arxiv.org/abs/2203.14343

[5] J. T. H. Smith, A. Warrington, and S. W. Linderman, "Simplified state space layers for sequence modeling," 2023. [Online]. Available: https://arxiv.org/abs/2208.04933

[6] K. Pfeuffer, M. J. Geiger, S. Prange, L. Mecke, D. Buschek, and F. Alt, "Behavioural biometrics in vr: Identifying people from body motion and relations in virtual reality," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.

[7] J. Liebers, M. Abdelaziz, L. Mecke, A. Saad, J. Auda, U. Gruenefeld, F. Alt, and S. Schneegass, "Understanding user identification in virtual reality through behavioral biometrics and the effect of body normalization," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–11.

[8] J. Liebers, C. Burschik, U. Gruenefeld, and S. Schneegass, "Exploring the stability of behavioral biometrics in virtual reality in a remote field study: Towards implicit and continuous user identification through body movements," in *Proceedings of the 29th ACM Symposium on Virtual Reality Software and Technology*, 2023, pp. 1–12.

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

[10] ——, "Combining real-world constraints on user behavior with deep neural networks for virtual reality (vr) biometrics," in *2022 IEEE conference on virtual reality and 3D user interfaces (VR)*. IEEE, 2022, pp. 409–418.

[11] L. Schach, C. Rack, R. P. McMahan, and M. E. Latoschik, "Motion-based user identification across xr and metaverse applications by deep classification and similarity learning," *arXiv preprint arXiv:2509.08539*, 2025.

[12] F. J. Ordóñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.

[13] H. Ismail Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P.-A. Muller, and F. Petitjean, "Inceptiontime: Finding alexnet for time series classification," *Data Mining and Knowledge Discovery*, vol. 34, no. 6, pp. 1936–1962, 2020.

[14] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396*, 2021.

[15] C. Rack, T. Fernando, M. Yalcin, A. Hotho, and M. E. Latoschik, "Who is alyx? a new behavioral biometric dataset for user identification in xr," *Frontiers in Virtual Reality*, vol. 4, p. 1272234, 2023.

[16] S. Somvanshi, M. M. Islam, M. S. Mimi, S. B. B. Polock, G. Chhetri, and S. Das, "From s4 to mamba: A comprehensive survey on structured state space models," 2025. [Online]. Available: https://arxiv.org/abs/2503.18970