Continuous ASL Fingerspelling Recognition from Wearable Accelerometers

Zhen Hou
Yeshiva University
New York, NY, USA

Yucheng Xie*
Yeshiva University
New York, NY, USA

Feng Li
Purdue University
Indianapolis, IN, USA

Honggang Wang
Yeshiva University
New York, NY, USA

***Abstract*—American Sign Language (ASL) fingerspelling recognition in public settings requires sensing that is privacy-preserving, portable, and unobtrusive. Wearable ring-mounted accelerometers are a promising modality for this setting, but existing systems remain difficult to scale because real wearable training data are scarce and current formulations do not support continuous, sentence-level, open-vocabulary recognition. We present *FingerSeq*, a framework for continuous, open-vocabulary ASL fingerspelling recognition from sparse wearable accelerometer input. FingerSeq uses large-scale video-derived virtual accelerometer supervision, formulates recognition as direct character-sequence decoding of continuous accelerometer streams, and adapts the resulting model to real wearable data through fine-tuning. On a large-scale public ASL fingerspelling dataset, FingerSeq achieves 32.76% CER with a sparse 5-point wearable configuration. After adaptation to real Tap Strap 2 accelerometer data, it achieves 14.13% CER and consistently outperforms CTC-based baselines. These results support ring-based accelerometer sensing as a promising direction for continuous fingerspelling recognition.**

***Index Terms*—ASL fingerspelling, wearable sensing, privacy-preserving sensing, continuous fingerspelling recognition**

I. Introduction

In American Sign Language (ASL), fingerspelling spells out words letter by letter through a sequence of hand gestures and is often used for words that may not have a standard sign, such as names or places [1]. Accurate recognition of fingerspelled words is especially important in public interactions, where misunderstanding essential information such as an address or medication name can disrupt communication entirely. However, many existing sensing approaches remain poorly suited to these settings: camera-based methods require users to actively position themselves within view of a camera and may raise privacy concerns in shared spaces [2], [3], while RF-based approaches [4], [5] depend on dedicated infrastructure and controlled environments. Wearable sensing is a more practical fit for everyday fingerspelling recognition because it does not depend on external infrastructure and better supports mobile use [6], [7].

Prior work shows the feasibility of wearable sensing for fingerspelling recognition [8]–[10]. However, wearable fingerspelling recognition in everyday use remains limited. First, real wearable fingerspelling training data are scarce because collecting them at scale is inherently difficult. In addition, reusable wearable fingerspelling datasets remain limited because many are not publicly available and existing datasets often differ in hardware setups and collection protocols [8], [9], [11]. Second, many existing systems rely on custom-designed wearable hardware [12], [13] or multi-device wearable setups [14], which highlights the need for lightweight, low-cost sensing configurations. Finally, existing methods are typically evaluated on isolated letter classification [15], [16] or word-level recognition [11], and many also rely on pre-segmented inputs or lexicon-constrained decoding [9], [11], [15], [16]. This makes them poorly suited to continuous, sentence-level, open-vocabulary fingerspelling recognition.

To address these challenges, we propose *FingerSeq*, a framework for continuous, open-vocabulary fingerspelling recognition that supports practical deployment with low-cost commercial wearables. To address the scarcity of real wearable training data, we propose a virtual accelerometer construction method based on video-derived hand landmark trajectories from large-scale public ASL video datasets. This provides large-scale supervision for robust sequence learning while preserving motion patterns that more closely reflect real human articulation than purely physical simulation. We then develop a direct character-sequence decoding approach that maps continuous, unsegmented wearable fingerspelling to sentence-level, open-vocabulary transcription without predefined lexicons. To support lightweight, low-cost everyday deployment, we use commercial ring-based accelerometer sensing with real-data adaptation, enabling effective recognition with as few as five finger-mounted sensors.

Our main contributions are:

- We present *FingerSeq*, to our knowledge the first system for continuous, open-vocabulary ASL fingerspelling recognition using low-cost commercial wearables.
- We develop the core methodology of FingerSeq, including video-to-virtual accelerometer construction for scalable supervision, direct character-sequence decoding for continuous open-vocabulary recognition, and virtual-to-real adaptation for deployment on commercial wearables.
- We evaluate FingerSeq on large-scale public ASL fingerspelling data and real wearable data, achieving 14.13% CER after virtual pretraining and real-data fine-tuning while consistently outperforming CTC-based baselines across both virtual and real wearable settings.

*Yucheng Xie is the corresponding author. Email: yucheng.xie@yu.edu.

II. Related Work

A. Sensing Approaches for Sign Language and Fingerspelling Recognition

Recognition of sign language and fingerspelling in public settings requires sensing approaches that remain practical for everyday use. Video-based methods include RGB-camera and depth-camera approaches for fingerspelling and sign language recognition [2], [3], [17]. In public settings, however, they are often impractical because users must either remain within view of the camera or actively position a camera-enabled device during communication. Always-on camera sensing can also raise privacy concerns in shared spaces because it may capture both users and bystanders. RF-based approaches include WiFi, mmWave, and RFID sensing for sign language and gesture recognition [4], [5], [18], [19]. They pose fewer privacy concerns but typically rely on external sensing infrastructure and constrained deployment setups, limiting mobility and practicality in public settings.

Wearable sensors offer a lightweight, privacy-preserving, and portable alternative for sign language and fingerspelling recognition. Existing examples include smartwatch- and EMG-based systems for sign language translation [14] and ring- or glove-based systems for fingerspelling recognition [8]–[10], [20]. These systems infer signed content from finger motion and orientation through acceleration and angular velocity patterns associated with handshape transitions. However, existing wearable approaches still face a fundamental data bottleneck. Real wearable fingerspelling datasets are difficult to scale, often differ across hardware setups and collection protocols, and provide limited support for large-scale continuous sequence learning [8]–[10], [20]. As a result, data scarcity remains a major barrier to practical wearable fingerspelling recognition.

B. Data Synthesis for Fingerspelling

Data synthesis has emerged as a practical strategy for addressing the data scarcity that limits wearable fingerspelling systems [21]. Representative methods differ in how they construct training signals for wearable recognition. Generative augmentation methods, including conditional GANs [22] and VAE-based augmentation [23], can increase the amount of available training data. However, because they remain tied to the distribution and coverage of the source datasets, they are less suited to generating the diverse training signals needed for wearable fingerspelling recognition. Physics-based simulators synthesize IMU readings from rigid skeletal kinematics, but this abstraction may not fully capture the soft-tissue dynamics and biomechanical coupling that shape real inertial signals [24], [25]. Language-to-motion synthesis approaches such as IMUGPT 2.0 [26] generate virtual IMU data by converting textual activity descriptions into motion sequences and then mapping those motions into wearable signals, but the generated motions may not preserve the fine-grained hand articulation needed for wearable fingerspelling recognition.

Within this design space, one particularly relevant direction for wearable fingerspelling is cross-modal synthesis from video-derived hand landmarks. It preserves the spatiotemporal structure of real human motion, provides per-joint landmark resolution, and can draw on large-scale public ASL corpora spanning hundreds of users and millions of annotated characters [2], [27]–[29]. At the hand level, systems such as Vi2IMU [15], ZeroNet [16], and SignRing [11] show that video-derived hand motion can support wearable-style recognition. However, limitations remain in how these systems formulate the recognition task. Vi2IMU and ZeroNet are formulated as isolated letter classification tasks, which assume pre-segmented units [15], [16]. As a result, they bypass the boundary detection, temporal alignment, and error accumulation challenges that arise in continuous sequence decoding. SignRing moves beyond isolated letters, but still predicts at the word level [11], which avoids open-ended character-level decoding over continuous fingerspelled streams. In addition, dictionary- or language-model-constrained post-processing restricts predictions to a predefined lexicon [11], [15], [16], making these approaches poorly suited to names and other unseen words that commonly appear in fingerspelling [1]. As a result, existing cross-modal approaches still fall short of continuous, sentence-level, open-vocabulary fingerspelling recognition.

III. Methodology

A. System Overview

FingerSeq comprises three components for continuous, open-vocabulary wearable fingerspelling recognition, as illustrated in Figure 1. These components are video-to-virtual accelerometer construction for scalable supervision, direct character-sequence decoding for continuous open-vocabulary recognition, and virtual-to-real adaptation for deployment on commercial wearables. During training, public ASL videos are first processed with hand landmark tracking, and virtual 3-axis accelerometer sequences are constructed from the resulting landmark trajectories at ring-aligned finger positions. These virtual accelerometer sequences are paired with ground-truth transcriptions to pretrain the sequence-to-sequence recognizer used for direct character-sequence decoding. The pretrained model is then fine-tuned on real wearable accelerometer data with augmentation for virtual-to-real adaptation. During inference, the adapted model maps continuous multichannel accelerometer streams directly to character sequences without explicit boundary detection or vocabulary constraints.

B. Video-to-Virtual Accelerometer Construction

Real wearable fingerspelling data are scarce and difficult to collect at scale. To obtain large-scale supervision aligned with sparse ring-based sensing, we derive virtual 3-axis accelerometer sequences from hand landmark trajectories extracted from the Google ASL Fingerspelling Recognition dataset [30]. This representation preserves the finger-wise temporal structure needed for wearable recognition while remaining aligned with sparse ring-based sensing. For each video frame t, we extract hand landmarks with MediaPipe Holistic [31] and estimate

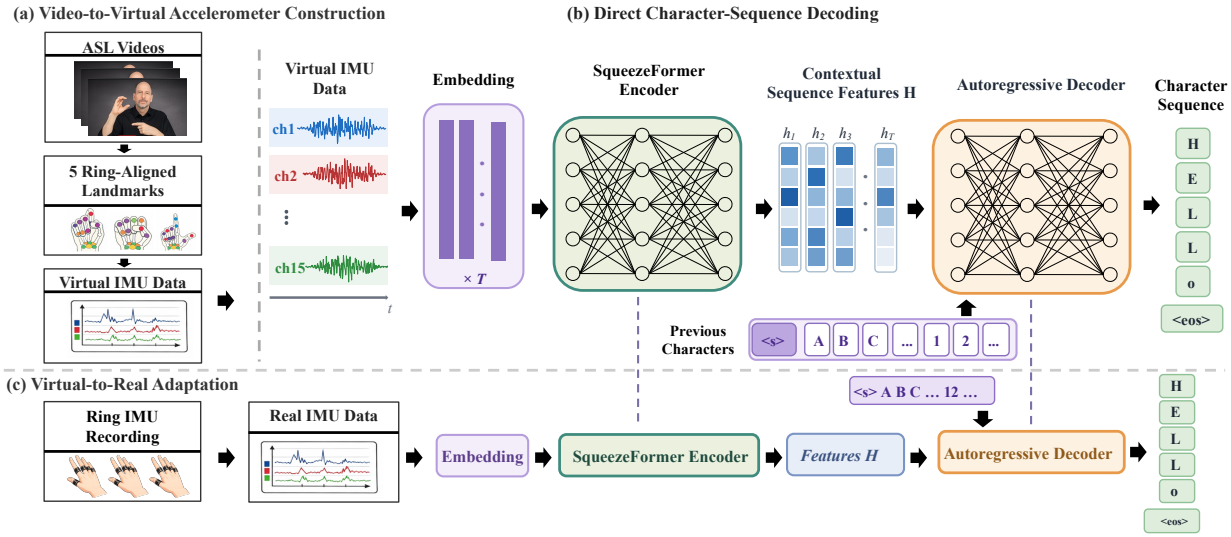


Fig. 1. Overview of FingerSeq. Video-derived hand landmarks are converted into virtual accelerometer sequences and paired with transcriptions to pretrain a sequence-to-sequence recognizer. The recognizer is then adapted to real wearable accelerometer data for continuous character-sequence decoding.

virtual acceleration from landmark trajectories by second-order finite differences:

$$\mathbf{a}_t^{(j)} = \frac{\mathbf{p}_{t+1}^{(j)} - 2\mathbf{p}_t^{(j)} + \mathbf{p}_{t-1}^{(j)}}{\Delta t^2}, \tag{1}$$

where $\mathbf{a}_t^{(j)}$ denotes the virtual 3-axis acceleration of landmark j at frame t, $\mathbf{p}_t^{(j)}$ is its 3D coordinate, and Δt is the video frame interval.

Because our real-device setup uses one sensing point per finger, the virtual representation must be reduced to the same sparse sensing layout for consistent pretraining and adaptation. We therefore retain five ring-aligned hand landmarks under the MediaPipe convention: thumb IP (3), index PIP (6), middle PIP (10), ring PIP (14), and pinky PIP (18). Concatenating their accelerations yields a 15-channel wearable proxy with the same one-sensor-per-finger layout as the real device. This representation allows large public ASL corpora to provide scalable wearable-aligned supervision while retaining realistic articulatory motion.

C. Direct Character-Sequence Decoding

Continuous sentence-level fingerspelling requires decoding an unsegmented accelerometer stream into a character sequence. Because letter boundaries are unreliable and target words cannot be restricted to a predefined lexicon, FingerSeq formulates recognition as direct sequence-to-sequence transduction from multichannel accelerometer signals to character sequences. This setting requires an encoder that can represent both local hand motion around individual letters and longer-range context across a continuous phrase. FingerSeq therefore uses a SqueezeFormer encoder [32], whose convolution and self-attention blocks are designed to combine local pattern modeling with long-range temporal dependencies.

The input stream is first mapped to frame-level embeddings and then processed by the SqueezeFormer encoder, which produces contextualized sequence representations $\mathbf{H}$. A 2-layer autoregressive Transformer decoder generates output characters from left to right by factorizing the conditional probability as

$$P(\mathbf{Y} \mid \mathbf{X}) = \prod_{i=1}^{L} P(y_i \mid y_{<i}, \mathbf{H}). \tag{2}$$

Here, $\mathbf{X}$ denotes the input accelerometer sequence, $\mathbf{Y} = (y_1, \ldots, y_L)$ the output character sequence of length L, and $y_{<i}$ the previously decoded characters. In our implementation, the encoder stacks six SqueezeFormer blocks with a hidden dimension of 144 and four attention heads, and the decoder uses the same hidden dimension. The same architecture is used across different input channel counts. This design avoids reliance on explicit framewise alignment and more naturally supports continuous character-sequence decoding when letter boundaries are unreliable and target words are not predefined.

D. Virtual-to-Real Adaptation

Although virtual accelerometer supervision provides scale, these signals remain only wearable-aligned proxies and do not fully capture the sampling characteristics, noise patterns, and physical placement variability of real ring-mounted accelerometers. For virtual-to-real adaptation, FingerSeq is first pretrained on large-scale video-derived virtual accelerometer sequences and then fine-tuned on real wearable recordings using the same cross-entropy sequence objective with label smoothing [33]. Fine-tuning is initialized from the virtual-pretrained checkpoint and optimized with AdamW (learning rate 1×10^{-3}, weight decay 0.05) under a cosine-annealing schedule with a batch size of 32. Because the real-device training set is limited, we further augment fine-tuning in two complementary ways. First, we perturb individual accelerometer sequences with random Gaussian noise, time stretching, and

per-axis amplitude scaling, using a noise standard deviation of 0.03 and sampling both time-stretching rates and scaling factors from $[0.8, 1.2]$. Crucially, a single time-warp and a single set of per-axis scale factors are sampled per example and applied to all five fingers, so augmentation does not break the cross-finger timing that fingerspelling depends on. These augmentations simulate sensor noise, variation in signing speed, and differences in ring fit or motion intensity. Second, we synthesize additional multi-letter training examples by concatenating isolated real-device letter recordings separated by short pauses of 3–10 frames. The corresponding character labels are concatenated in the same order. Importantly, these concatenated sequences are synthetic augmentation rather than natural phrase-level recordings, so they do not reproduce the full coarticulation, timing variability, and transition dynamics of continuous fingerspelling.

IV. Performance Evaluation

A. Experimental Setup

Settings. We evaluate FingerSeq in two experimental settings: a virtual setting based on video-derived accelerometer supervision and a real wearable setting based on physical accelerometer input collected with a commercial ring-based device. The virtual setting is used to evaluate sentence-level recognition, open-vocabulary generalization, and sensor-density effects, while the real wearable setting is used to evaluate both training from scratch and virtual-to-real adaptation.

Dataset. For the virtual setting, we use the Google ASL Fingerspelling Recognition dataset [30], which contains 94 participants and approximately 67,000 continuous fingerspelling videos annotated with character sequences. We adopt a within-person phrase-level split, so training and test data come from the same signers but contain disjoint phrase instances. We use the same 15-channel wearable-aligned virtual accelerometer representation described in Section III, constructed from five ring-aligned hand landmarks extracted with MediaPipe Holistic.

Device. Figure 2 shows the Tap Strap 2 [34] setup used in our real wearable setting. The device is worn across the five fingers, with one ring positioned on each finger to match the virtual sensing layout. Each ring provides a 3-axis accelerometer sampled at 200 Hz, yielding 15 channels in total. The real-device dataset contains 943 training samples and 63 held-out test samples. The recordings consist of short fingerspelled phrases representative of everyday communication scenarios.

Evaluation Metrics. We report **Character Error Rate (CER)**, computed as the normalized Damerau–Levenshtein distance between the predicted and ground-truth character sequences, and **Exact Match Rate (EMR)**, the percentage of test phrases whose predicted character sequence exactly matches the ground truth. CER captures fine-grained transcription errors, whereas EMR reflects end-to-end phrase correctness.

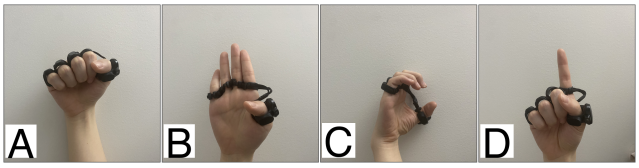
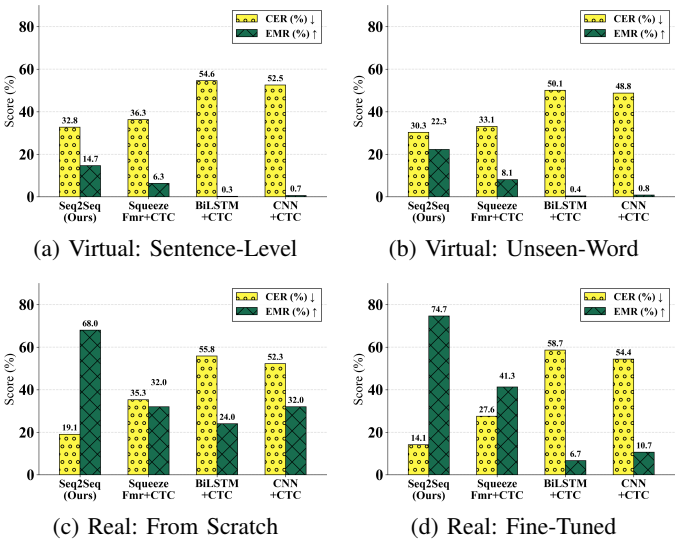


Fig. 2. The Tap Strap 2 worn on the dominant hand, shown performing four ASL fingerspelling gestures: A, B, C, D. The rings are positioned near the PIP region of the four non-thumb fingers and the corresponding IP region of the thumb. Each ring contains a 3-axis accelerometer, yielding 15 channels (5 fingers $\times$ 3 axes) at 200 Hz.



(a) Virtual: Sentence-Level (b) Virtual: Unseen-Word

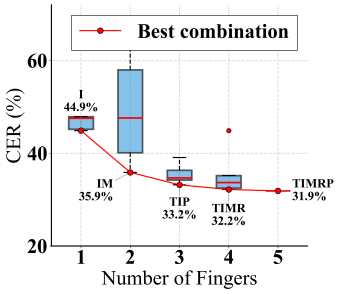
(c) Real: From Scratch (d) Real: Fine-Tuned

Fig. 3. Recognition performance on the wearable configuration. FingerSeq achieves the strongest overall performance across all settings.

B. Comparison with Baseline Methods

To evaluate FingerSeq against common baselines for continuous unsegmented recognition, we compare it with three models that use the connectionist temporal classification (CTC) objective [35], [36]. These baselines differ only in encoder architecture and are implemented as SqueezeFormer+CTC, BiLSTM+CTC, and CNN+CTC. All models use the same input representation, front-end feature extractor, and 5-point wearable configuration, and are trained for 80 epochs with AdamW at a learning rate of 4.5×10^{-3} on the same continuous, unsegmented fingerspelling sequences.

Overall Performance. FingerSeq consistently outperforms the strongest CTC baseline across both virtual and real wearable settings. In the virtual setting (Fig. 3(a)), both training and testing use video-derived virtual accelerometer data, and FingerSeq achieves the lowest CER (32.76%) and highest EMR (14.67%), compared with 36.32% CER and 6.35% EMR for SqueezeFormer+CTC. In the real wearable setting (Fig. 3(c)), both training and testing use real accelerometer data, and all models are trained from scratch. Under this setting, FingerSeq again performs substantially better, reaching 19.08% CER and 68.00% EMR versus 35.34% CER and 32.00% EMR. Most importantly, virtual pretraining followed by real-data



(a) CER vs. Finger Count

Config	ΔCER	Config	ΔCER
−T	3.3	−T−I	7.2
−I	1.8	−T−M	4.8
−M	0.6	−T−R	4.0
−R	13.0	−T−P	4.6
−P	0.3	−I−M	3.1
		−I−R	2.3
		−I−P	2.5
		−M−R	1.3
		−M−P	2.5
		−R−P	1.8

(b) Finger Removal Impact

Fig. 4. Finger configuration analysis in the virtual setting. (a) CER distribution across all finger subsets grouped by the number of instrumented fingers. (b) CER increase (percentage points) after removing one or two fingers from the full five-finger baseline (CER = 31.9%). Config denotes the removed fingers, and ΔCER denotes the resulting CER increase.

fine-tuning further improves FingerSeq to 14.13% CER and 74.67% EMR (Fig. 3(d)). Although SqueezeFormer+CTC also benefits from fine-tuning, improving to 27.56% CER and 41.33% EMR, FingerSeq retains a substantial advantage after adaptation.

Open-Vocabulary Generalization. To evaluate character-level generalization beyond memorized training phrases, we report performance on the subset of virtual test phrases containing at least one out-of-vocabulary word, which we denote as *unseen* phrases. Figure 3(b) shows that FingerSeq achieves 30.32% CER and 22.26% EMR on unseen phrases, compared with 33.09% CER and 8.12% EMR for SqueezeFormer+CTC. These results indicate that FingerSeq maintains stronger character-level decoding on unseen phrases.

C. Finger Configuration Analysis

Beyond the main 5-point wearable configuration, we further analyze recognition performance across different finger configurations. To this end, we evaluate FingerSeq under all 31 finger-subset configurations (i.e., all possible subsets of five fingers) on the virtual setting. Figure 4(a) shows that CER drops rapidly as the number of sensors increases from one to three, but the gains become marginal beyond that point. The best single-finger configuration uses the index finger (I, 44.9%), while the best two-finger configuration uses the index and middle fingers (IM, 35.9%), yielding a 9.0 percentage point improvement. With three sensors, the best subset is TIP (33.2%), which comes within 1.3 percentage points of the full five-finger baseline (TIMRP, 31.9%).

Figure 4(b) further measures the CER increase when removing each finger or finger pair from the full five-finger configuration. Among single-finger removals, dropping the ring finger causes by far the largest degradation (+13.0 percentage points in CER), whereas removing the pinky has the smallest effect (+0.3). This indicates that the contributions of individual fingers are highly uneven under the full sensing configuration. Among two-finger removals, removing the thumb–index pair causes the largest degradation (+7.2), further underscoring that sensor placement matters as much as sensor count. In contrast, removing the middle–ring pair causes the smallest degradation (+1.3), corresponding to the thumb–index–pinky (TIP) configuration. This result identifies TIP (33.2%) as a practical three-finger configuration: it uses 40% fewer sensors while remaining within 1.3 percentage points of the full five-finger baseline.

V. Discussion

The findings of this study have several implications for wearable fingerspelling recognition. First, large-scale virtual accelerometer supervision provides useful training signal for continuous wearable fingerspelling recognition. This supports the use of video-derived virtual accelerometer data as a scalable pretraining signal rather than a fully real-data collection replacement. Second, FingerSeq consistently outperforms CTC-based baselines both in the virtual setting and after adaptation to real Tap Strap 2 data, indicating that the advantage of direct character-sequence decoding carries over to physical accelerometer input despite the domain gap between virtual and real sensing. Third, the sensor-density analysis shows that recognition accuracy does not scale linearly with sensor count alone. The strongest trade-off emerges around the thumb–index–middle configuration, while adding remaining fingers yields smaller gains. This suggests that practical wearable designs may not require full five-finger sensing, highlighting the importance of sensor placement over sheer sensor count. These results motivate future work on richer motion cues, such as gyroscope signals, to determine whether additional sensing modalities can further reduce sensor-count or placement constraints.

From a privacy and security perspective, accelerometer sensing offers a practical advantage: it avoids recording speech, video, or nearby bystanders and can support local inference on wearable or paired devices. In this work, we use video-derived virtual accelerometer supervision and a real Tap Strap 2 dataset to study whether continuous fingerspelling recognition is feasible under sparse commercial wearable sensing. This setting allows us to isolate the virtual-to-real adaptation problem and analyze sparse wearable sensing under controlled conditions. Building on these results, our next step is to evaluate FingerSeq across more users and devices, characterize robustness in less controlled settings, and develop formal privacy and security analyses.

VI. Conclusion

In this paper, we presented FingerSeq, a framework for continuous, open-vocabulary ASL fingerspelling recognition from wearable accelerometer input that supports deployment on low-cost commercial wearables. By combining video-derived virtual accelerometer supervision, direct character-sequence decoding, and fine-tuning on real wearable data, FingerSeq mitigates the scarcity of real wearable training data while supporting continuous, open-vocabulary recognition without lexicon-constrained decoding. FingerSeq consistently outperforms CTC-based baselines in both the virtual and real wearable settings, and the sensor-density analysis further

suggests that sparse ring-based accelerometer sensing is a promising direction for continuous fingerspelling recognition.

ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation under Grant Nos. 2428595, 2436733, and 2431725, and by the Enhancing Faculty Research in STEM and Health Grant from the Katz School of Science and Health, Yeshiva University.

REFERENCES

[1] C. Padden and D. Gunsauls, "How the alphabet came to be used in a sign language," *Sign Language Studies*, vol. 4, no. 1, pp. 10–33, 2003.

[2] B. Shi, A. Martinez Del Rio, J. Keane, J. Michaux, D. Brentari, G. Shakhnarovich, and K. Livescu, "Fingerspelling recognition in the wild with iterative visual attention," in *Proc. IEEE/CVF ICCV*, 2019, pp. 5400–5409.

[3] N. C. Camgöz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in *Proc. IEEE/CVF CVPR*, 2018, pp. 7784–7793.

[4] Y. Ma, G. Zhou, S. Wang, H. Zhao, and W. Jung, "SignFi: Sign language recognition using WiFi," in *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. (IMWUT)*, vol. 2, no. 1, pp. 1–21, 2018.

[5] P. S. Santhalingam *et al.*, "mmASL: Environment-independent ASL gesture recognition using 60 GHz millimeter-wave signals," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 1, art. 26, 2020, doi:10.1145/3381010.

[6] K. Kudrinko, E. Flavin, X. Zhu, and Q. Li, "Wearable sensor-based sign language recognition: A comprehensive review," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 82–97, 2021, doi:10.1109/RBME.2020.3019769.

[7] R. Tchantchane, H. Zhou, S. Zhang, and G. Alici, "A review of hand gesture recognition systems based on noninvasive wearable sensors," *Adv. Intell. Syst.*, vol. 5, no. 10, 2023, doi:10.1002/aisy.202300207.

[8] D. Martin *et al.*, "FingerSpeller: Camera-Free Text Entry Using Smart Rings for American Sign Language Fingerspelling Recognition," in *Proc. 25th Int. ACM SIGACCESS Conf. Computers and Accessibility (ASSETS '23)*, 2023, doi:10.1145/3597638.3614491.

[9] H. Lim *et al.*, "SpellRing: Recognizing Continuous Fingerspelling in American Sign Language using a Ring," in *Proc. ACM CHI*, 2025, doi:10.1145/3706598.3713721.

[10] Y. Gu *et al.*, "American Sign Language Alphabet Recognition Using Inertial Motion Capture System with Deep Learning," *Inventions*, vol. 7, no. 4, p. 112, 2022, doi:10.3390/inventions7040112.

[11] J. Li et al., "SignRing: Continuous American Sign Language recognition using IMU rings and virtual IMU data," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 7, no. 3, 2023.

[12] K. Suri and R. Gupta, "Convolutional neural network array for sign language recognition using wearable IMUs," in *Proc. 6th Int. Conf. Signal Process. Integr. Netw. (SPIN)*, 2019.

[13] K. Suri and R. Gupta, "Continuous sign language recognition from wearable IMUs using deep capsule networks and game theory," *Comput. Electr. Eng.*, vol. 78, pp. 493–503, 2019.

[14] Q. Zhang, J. Jing, D. Wang, and R. Zhao, "WearSign: Pushing the limit of sign language translation using inertial and EMG wearables," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 6, no. 1, 2022.

[15] P. S. Santhalingam *et al.*, "Synthetic Smartwatch IMU Data Generation from In-the-wild ASL Videos," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 7, no. 2, art. 74, 2023, doi:10.1145/3596261.

[16] Y. Liu, S. Zhang, and M. Gowda, "When Video meets Inertial Sensors: Zero-shot Domain Adaptation for Finger Motion Analytics with Inertial Sensors," in *Proc. ACM/IEEE IoTDI*, 2021, pp. 182–194, doi:10.1145/3450268.3453537. (Best Paper Award)

[17] T.-W. Chong and B.-G. Lee, "American Sign Language Recognition Using Leap Motion Controller with Machine Learning Approach," *Sensors*, vol. 18, no. 10, p. 3554, 2018, doi:10.3390/s18103554.

[18] H. Zhang *et al.*, "RF-Sign: Position-independent sign language recognition using passive RFID tags," *IEEE Internet Things J.*, vol. 11, no. 5, pp. 9056–9071, 2024, doi:10.1109/JIOT.2023.3322228.

[19] H. Xu, Y. Zhang, Z. Yang, H. Yan, and X. Wang, "RF-CSign: A Chinese sign language recognition system based on large kernel convolution and normalization-based attention," *IEEE Access*, vol. 11, pp. 133767–133780, 2023, doi:10.1109/ACCESS.2023.3333036.

[20] Y. Liu, F. Jiang, and M. Gowda, "Finger Gesture Tracking for Interactive Applications: A Pilot Study with Sign Languages," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 3, art. 112, 2020, doi:10.1145/3414117.

[21] A. Kamboj and M. Do, "A survey of IMU based cross-modal transfer learning in human activity recognition," arXiv:2403.15444, 2024.

[22] A. Hakim *et al.*, "cGAN-based high dimensional IMU sensor data generation for enhanced human activity recognition in therapeutic activities," *Biomed. Signal Process. Control*, vol. 99, 2025, doi:10.1016/j.bspc.2024.106903.

[23] Z. Li, C. Yu, C. Liang, and Y. Shi, "PoseAugment: Generative human pose data augmentation with physical plausibility for IMU-based motion capture," in *Proc. ECCV*, 2024.

[24] M. Muehlhausen *et al.*, "WIMUSim: Simulating realistic variabilities in wearable IMUs for human activity recognition," *Front. Comput. Sci.*, vol. 3, 2025, doi:10.3389/fcomp.2025.1514933.

[25] Y. Hao, X. Lou, B. Wang, and R. Zheng, "CROMOSim: A Deep Learning-Based Cross-Modality Inertial Measurement Simulator," *IEEE Trans. Mobile Comput.*, vol. 23, no. 1, pp. 302–312, 2024, doi:10.1109/TMC.2022.3230370.

[26] Z. Leng *et al.*, "IMUGPT 2.0: Language-Based Cross Modality Transfer for Sensor-Based Human Activity Recognition," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 8, no. 3, 2024, doi:10.1145/3678545.

[27] B. Shi, K. Livescu, and D. Brentari, "Fingerspelling detection in American Sign Language," in *Proc. IEEE/CVF CVPR*, 2021, pp. 4166–4175.

[28] A. Duarte *et al.*, "How2Sign: A large-scale multimodal dataset for continuous American Sign Language," in *Proc. IEEE/CVF CVPR*, 2021, pp. 2735–2744.

[29] T. Tavella *et al.*, "FSboard: Over 3 million characters of ASL fingerspelling collected via smartphones," arXiv:2407.15806, 2024.

[30] Kaggle, "American Sign Language Fingerspelling Recognition," https://www.kaggle.com/competitions/asl-fingerspelling, accessed Apr. 19, 2026.

[31] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hao, F. Liu, J. Fan, L. Wang, M. Grundmann, and V. Bazarevsky, "MediaPipe: A framework for building perception pipelines," 2019, arXiv:1906.08172.

[32] S. Kim, A. Gholami, A. Shaw, N. Lee, K. Mangalam, J. Malik, M. W. Mahoney, and K. Keutzer, "Squeezeformer: An efficient transformer for automatic speech recognition," in *Proc. NeurIPS*, 2022.

[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE/CVF CVPR*, 2016, pp. 2818–2826.

[34] Tap Systems, Inc., "Tap Strap 2," https://shop.tapwithus.com/products/tapwithus-tap-strap-2-wearable-keyboard-mouse-air-gesture-controller, accessed Apr. 19, 2026.

[35] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006, pp. 369–376.

[36] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/Attention architecture for end-to-end speech recognition," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1240–1253, 2017.