ASLRing: American Sign Language Recognition with Meta-Learning on Wearables

Hao Zhou
Pennsylvania State University

Taiting Lu†
Pennsylvania State University

Kenneth DeHaan
Gallaudet University

Mahanth Gowda
Pennsylvania State University

†Co-primary

***Abstract*—Sign Language is widely used by over 500 million Deaf and hard of hearing (DHH) individuals in their daily lives. While prior works made notable efforts to show the feasibility of recognizing signs with various sensing modalities both from the wireless and wearable domains, they recruited sign language learners for validation. Based on our interactions with native sign language users, we found that signal diversity hinders the generalization of users (e.g., users from different backgrounds interpret signs differently, and native users have complex articulated signs), thus resulting in recognition difficulty. While multiple solutions (e.g., increasing diversity of data, harvesting virtual data from sign videos) are possible, we propose *ASLRing* that addresses the sign language recognition problem from a meta-learning perspective by learning an inherent knowledge about diverse spaces of signs for fast adaptation. *ASLRing* bypasses expensive data collection process and avoids the limitation of leveraging virtual data from sign videos (e.g., occlusions, overexposure, low-resolution). To validate *ASLRing*, instead of recruiting learners, we conducted a comprehensive user study with a database with 1080 sentences generated by a vocabulary size of 1057 from 14 native sign language users and achieved a 26.9% word error rate, and we also validated *ASLRing* in diverse settings[1].**

I. Introduction

Sign language is a natural language that serves primarily in deaf communities. According to the World Health Organization (WHO), the population of the Deaf and Hard of Hearing (DHH) individuals reaches 10 million in the USA and ≈500 million globally [1], [39]. By 2050, over 700 million people will suffer from hearing loss [5]. Sign language is their primary means of communication, while most hearing people with no sign language experience always have difficulty understanding the sign language performed by Deaf people. In recent years, it has been brought to the attention of the entire society that it is crucial and necessary to bridge the communication barrier between DHH individuals and hearing people.

Past research in ubiquitous computing and computer vision communities has explored different ways to facilitate communications between Deaf people and hearing people. There has been much recent work on addressing sign language recognition (SLR) problems via different wireless or wearable sensing methods such as WiFi [37], EMG sensor [56], and motion sensor [33]. Although great efforts have been made, simplifications such as reducing the complexity of sign sentences [33], and only capturing signs from one hand (both hands are important in sign languages) [33], [56], in the systems also hinder wide adoption. Moreover, most of the works evaluated their systems only on sign learners and students because of the difficulty of recruiting native ASL signers. Through our extended interaction with native sign language users, we reveal several insights on the challenging sign language recognition task. Firstly, it is trivial to imagine that for a visual language, signs from native users will inevitably be more articulated than beginners, thus imposing challenges in recognition. Secondly, like speech diversity due to regions, sign languages are also diverse across different, cultures, and communities. In contrast to these systems, we conducted a comprehensive user study involving native ASL signers to investigate and reveal the inherent complexities and difficulties associated with SLR tasks. Also, while cameras-based methods [13], [38] can be another alternative to wireless or wearable sensing, sensors-based methods are robust to lighting and resolution and have fewer privacy concerns.

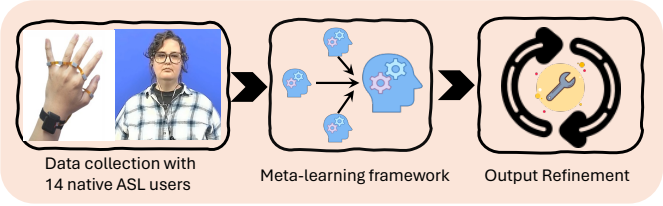


Fig. 1: Overall flow of *ASLRing*. We employ meta-learning training strategies and propose a simple yet effective output refinement algorithm. In a nutshell, *ASLRing* achieves a 26.9% word error rate across 14 native American Sign Language Users.

To address the challenges, prior works' solution focuses on increasing the diversity of the training data (e.g., acquiring more data from various users), or harvesting virtual data from available sources from other domains (e.g., virtual inertial data from sign videos). While these solutions can improve recognition performance, collecting data with native users is expensive, and virtual data inevitably inherit issues such as occlusion, and overexposure from sign videos. In contrast, we adopt the principles of meta-learning [8], a new learning paradigm that encapsulates the idea of training models to rapidly adapt to new tasks using minimal data. Instead of focusing on mastering a single task, as in traditional machine learning,

[1]Project page: https://www.cse.psu.edu/~mkg31/projects/aslring/

meta-learning emphasizes acquiring a broader learning strategy that can be applied across various tasks. Briefly, we find it a good fit for sign language recognition tasks with diversity from various factors (e.g., users, signing styles, dialects). Our key insight is that instead of treating all data equally like in traditional sign language recognition systems, we define each sign sentence signed by various users as a task in meta-learning. By doing so, we build a system that decouples signal variations from users. Moreover, by developing an inherent "intuition" across diverse spaces of sign sentences during training, *ASLRing* acquires the ability to generalize on unseen sentences with learned knowledge of the sentence space.

We also observe a *partial correctness* behavior. When a different input signal (e.g., temporal shifted, masked, noise-injected versions) is given, the deep learning model only captures partial answers because the model has learned certain patterns from the input effectively. Still, it hasn't completely generalized for the entirety of the inputs due to highly articulated signals from native users, leading to inconsistent or fragmented outputs on variations. To refine the prediction, we propose a simple yet effective refinement based on N-gram models [12]. In short, we combine results from diverse input signals and predict the next word with N-gram models.

Combining together, we propose *ASLRing*, a meta-learning-aided system for sign language recognition with output refinement. Fig. 1 depicts the high-level overview of the system where wearable sensors are used in *ASLRing* to capture signs by native users, and the sensory information is processed by our meta-learning framework and the result is further refined for later usage such as translating into English for ordering food, asking direction, etc.

An extensive native ASL user study with a group of Deaf native ASL signers is involved to validate the performance of *ASLRing* with a diverse database with 1080 sentences generated by a vocabulary size of 1057 that covers topics like sports, life, education, etc, from 14 Deaf users. In a nutshell, *ASLRing* achieves 26.9% word error rate (WER) across users. A more challenging case, unseen sentences, is also studied and we show *ASLRing* achieves ≈40% WER. Furthermore, the accuracy is also consistent across variations in signing speed, sensor wearing positions, and dialects/accents, demonstrating robustness.

In summary, we enumerate our contributions below. ❶ We reveal our insights on sign language recognition tasks based on our extended interaction with native sign language users. ❷ We propose a sign language recognition framework to address the diversity issues in signed languages by involving meta-learning training strategies. We define sentences as tasks to learn an inherent knowledge of diverse spaces of sign sentences. ❸ An extensive study with 14 native sign language users is conducted and an average 26.9% word error rate is achieved across various users. *ASLRing* is also validated in diverse settings including signing speed, etc.

II. Background

In this section, we briefly introduce the characteristics of inertial measurement units (IMU) and a broader impression of American Sign Language and American Sign Language Recognition Task.

A. IMU Data Introduction.

IMU sensors are cheap and widely used in wearable applications in sports analytics [19], AR/VR [47], human activity recognition [62], etc. An IMU sensor consists of three primary sensors: an accelerometer, a gyroscope, and a magnetometer. An accelerometer sensor measures the net acceleration including translational acceleration and the constant gravitational force vector. A magnetometer sensor measures the spatial orientation of the Earth's magnetic field, whereas a gyroscope sensor measures angular velocity and rotational motion. The accelerometer measurements can be converted from *local frame* of reference to a *global frame* of reference [61], by computing the orientation of the sensor. While prior research may exhibit susceptibility to environmental magnetic interference, *ASLRing* demonstrates immunity to such interference. *ASLRing* employs opportunistic error compensation techniques derived from A3 [61], to periodically reset drifts in gyroscope integration and minimize the effects of magnetic interference and motion artifacts. Though during data collection, we encountered environmental magnetic interference from ferromagnetic objects such as metallic framing, furniture, etc, we do not observe drift in errors in *ASLRing* as validated in Sec. VII.

B. Basics of American Sign Language

Sign Languages are visual languages that are composed of complex gestures and grammar. We begin with a brief overview of American Sign Language (ASL) to give a broader impression of sign language.

Signs and Gestures: Like any other sign language that uses gestures instead of sound for communication, ASL is also a form of natural language with its own grammar and lexicon. While approximately 137 sign languages are used by millions of Deaf people worldwide, ASL benefits Deaf people who mainly live in the USA and parts of Canada. In sign languages, most signs are a sequence of gestures from both hands (with one dominant hand and fingers involved). Fig. 2a shows the hand poses for finger spellings - A, S, and L. Fig. 2b and Fig. 2c show hand motions involved in signing "bike" and "learn".

(a) ASL Finger-spelling **(b)** ASL sign: Bike **(c)** ASL sign: Learn

Fig. 2: Examples of hand gestures and motion in ASL.

As shown in the figures, while the dominant hand performs a sequence of motion, the other hand (so-called

TABLE I: ASL sentence vs English sentence

ASL Sentence	English Sentence
AGE YOU	How old are you?
ME ARRIVE LATE SORRY	Sorry, I am late.
ME DRINKalc WINE RED	I'm drinking red wine.
LEARN SIGN WANT	I want to learn sign language.

non-dominant hand) is used to complement the dominant hand gestures to make the signs rich and meaningful. Similarly, facial expressions can also complement the signs. The entire signing motion including both hands and/or facial expressions is denoted as *gloss* in linguistic terms, which is an intermediate representation between sign languages and spoken language such as English. We use *gloss* to represent the connection between sign languages and spoken languages, and *sign* to represent the actual sequence of hand motion.

ASL Grammar: Sentences are mainly in the following format: *Subject-Verb-Object* [9]. The order of glosses can change when the context of the topic is established first through topicalization [41], leading to the following format: *Object, Subject-Verb*. Table I shows a few examples of ASL sentences and English translations. Micro-pauses happen at the end of sentences, which makes segmenting individual signs complex, especially for those who are native to the language.

American Sign Language Recognition: To build a connection between Deaf and hearing communities, researchers are devoted to developing systems that can translate *signs* into *glosses*. This step is named sign language recognition (SLR) as shown in Fig. 3. Furthermore, based on the complexity of segmenting individual signs from a sequence of hand motions, the SLR task is further divided into two problems: Isolated Sign Language Recognition (ISLR) and Continuous Sign Language Recognition (CSLR). While ISLR requires manually segmented signs, which requires extreme effort, CSLR only requires glosses in the form of a sentence. Given that CSLR is more reflective of everyday communication for the Deaf community, *ASLRing* concentrates on CSLR.

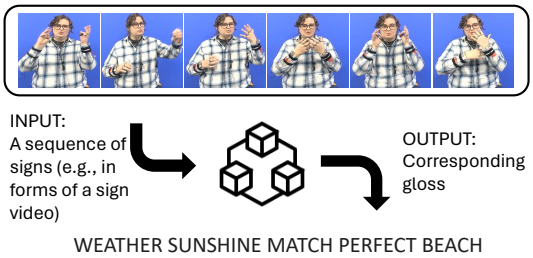


Fig. 3: Illustration of (Continuous) Sign Language Recognition. A recognition system takes a sequence of signs as input and outputs the corresponding gloss.

III. Learning From Native ASL Users

Why is sign language hard to recognize? To answer this question, we share our interactions with native ASL users, and summarize the observations that direct *ASLRing*.

Observation 1: *Signs are replaceable with a different sign and omittable without disturbing the meaning.*

It is easy to show that in any spoken language, there are always alternatives to express the same meaning. This holds true for sign languages. For example, "I love you" is semantically the same as "I fall in love with you". The corresponding glosses are totally different: "LOVEchest" and "ME FALL-IN-LOVE YOU". Fig. 4 demonstrates the signal difference between these two sentences in the form of IMU data. Native ASL users sign diversely, given their ability to interpret a sentence in various ways. However, when researchers, especially those lacking domain expertise, design recognition systems, the inconsistency between signs and labels can lead to systems that might not adapt well to these variations, which are an inherent part of communication within the Deaf community. Therefore, placing emphasis on semantic interpretation rather than solely focusing on the one-to-one mapping between signs and labels is crucial to address the challenge.

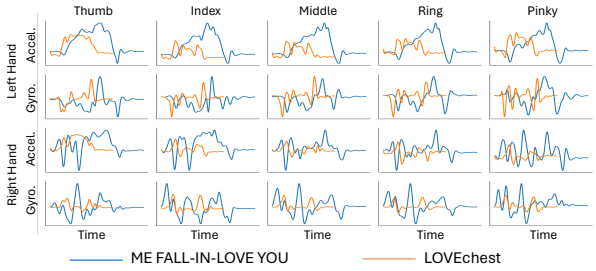


Fig. 4: Illustration of Observation 1 where two sentences are interchangeable to use because of the similar semantic meaning. However, the signals are quite different, indicating the challenge of recognizing signs.

Observation 2: *Sign language users also have "dialect" due to communities, cultures, styles, etc.*

Dialects in sign languages can be as varied and nuanced as spoken languages. Different regions, cultures, and communities may develop unique signing styles or variations on common signs due to a variety of factors. As a form of visual language, dialect in sign languages is mainly represented by various factors such as signing orders, hand shapes, and movement paths. As told by our users, 1) Asian Americans often structure their signs differently than White sign language users when signing for the same sentence; 2) Some users might use both hands to perform sign completely, while others tend to finish sign with only one hand without disturbing the meaning. Fig. 5 demonstrates such an example of signing by users with different styles, and it is obvious to see the signals vary. Moreover, native signers will have more complex co-articulated signs than sign language learners, imposing another challenge for the recognition task.

In a nutshell, these observations highlight the challenges faced when trying to recognize the glosses from diverse signal inputs. While some studies emphasize the frequency of specific sign occurrences, co-articulated signs [56]. Various signing styles make directly applying their methods problematic. Although expanding data diversity is a plausible solution, the process of gather-

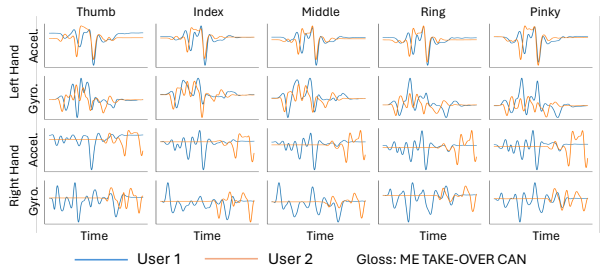


Fig. 5: Illustration of Observation 2 where one sentence is signed by native ASL users with different signing styles. It clearly notes that signals from the users vary (e.g., when signing for the same sentence, User 2's right hand barely moves, while User 1's right hand also complements the signs, demonstrating the various signing habits they developed over years).

ing large-scale sign language data is undoubtedly labor-intensive. Some researchers [29], [33] have tried sourcing virtual data from existing sign language datasets. However, virtual IMU data come with their own set of challenges, including occlusion, motion blur, and overexposure. These issues compromise the quality of the virtual data, affecting the efficacy of systems built on them.

IV. *ASLRing*

A. Why Use Meta-learning?

Unlike previous approaches, *ASLRing* circumvents the costly data acquisition phase and sidesteps the constraints of using virtual data from sign videos (such as occlusions, overexposure, and low resolution) by incorporating meta-learning into the process. *Why use meta-learning?* We begin with the example of dog classification from images. A single image can contain various noise factors: other animals, different backgrounds, multiple dog poses, or distinct breeds. Each of these elements can potentially affect the ability to accurately identify a dog. This situation parallels the challenges faced in the SLR domain: for a given sentence, diverse signing styles and interchangeable signs can introduce complexity to IMU signals, thereby compromising sentence recognition precision. In *ASLRing*, we argue that the successful principles of meta-learning in image classification can similarly be harnessed for sign language recognition. In contrast to supervised learning, which trains systems based on direct one-to-one correspondence, meta-learning employs training techniques that enable the recognition system to acquire comprehensive knowledge spanning various task spaces, as illustrated in Fig. 6a. This approach equips the system with the capability to handle other challenges such as signal variations (Sec. III) using limited data and achieving better generalization on unseen samples. Next, we elaborate on the details of *ASLRing* on how we define meta-learning concepts under the hood of sign language recognition.

B. Learning via Meta-learning

Dataset: Given a dataset of IMU signals, our goal is to recognize the signals into corresponding glosses. Let $\mathcal{D}$ be the dataset containing pairs of IMU signals and their corresponding glosses, i.e., $\mathcal{D} = \{(x_i^j, y_i)\}$, where x_i^j is the j^{th} varied IMU signal for sentence i and y_i is the corresponding gloss label.

Task: In the realm of meta-learning applied to sign language recognition, defining *tasks* poses a unique challenge. While in image classification, tasks can be straightforwardly categorized (e.g., each class as a distinct task), the definition is less clear for sign language recognition. One might consider defining each gloss as a task, but this demands well-segmented signs, which is a complex task for native users. In *ASLRing*, we propose a novel task formulation by designating sign sentence i (denoted as x_i) as task i. This definition offers dual advantages: 1) It separates the variations introduced by users in input signals for a consistent sentence. 2) A model trained at the sentence level can accumulate wide-ranging semantic knowledge across all task (sentence) domains, thereby enhancing its capability to make predictions on unseen sentences. With such a definition, we can form a set of tasks $\mathcal{T}$, where each task $T \in \mathcal{T}$ corresponds to mapping one sentence (i.e., varied signal inputs due to factors such as signing styles, replaceable signs) into its corresponding glosses.

Support Set and Query Set: For each task T, we have a support set $\mathcal{S}_T$ and a query set $\mathcal{Q}_T$. For example, for sentence i, the corresponding support set $\mathcal{S}_T = \{(x_i^s, y_i)\}$ and the query set $\mathcal{Q}_T = \{(x_i^q, y_i)\}$. Note that the sum of s and q denotes the total number of varied IMU signals (e.g., sentence i was signed by different ASL users) for sentence i.

Objectives in Meta-learning: ❶ *Task-level Learning*: Given a task T, we aim to learn a model (f_θ, parameterized by θ) that can quickly adapt to new, previously unseen variations using a small number of IMU signals from support set $\mathcal{S}_T$. In *ASLRing*, we adopt the classic Model-Agnostic Meta-Learning (MAML) [8] and the process can be represented as:

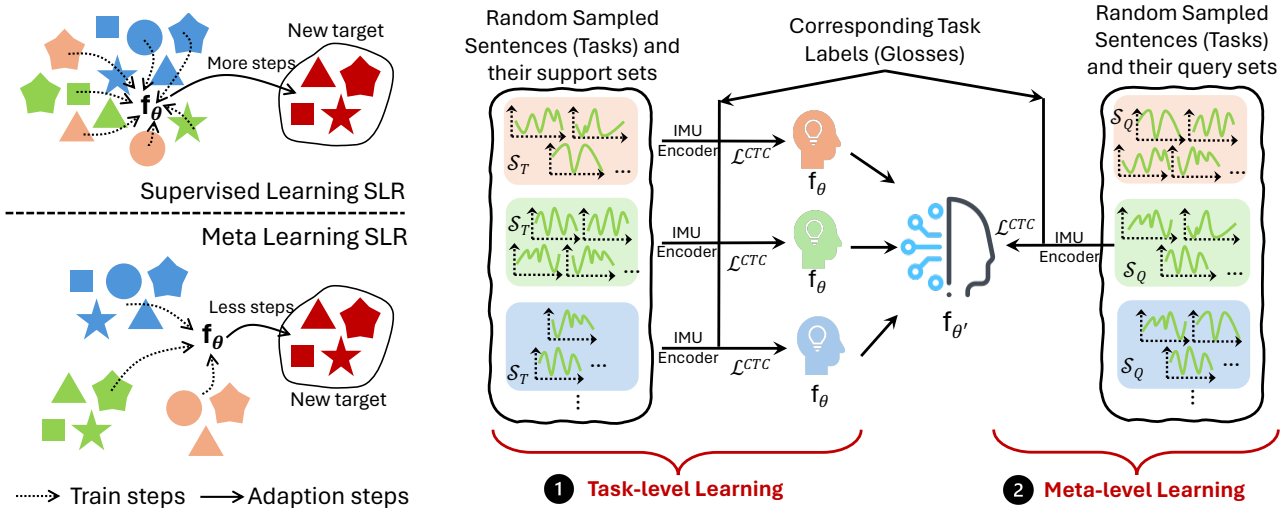
$$\theta' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{S}_T}(f_\theta), \tag{1}$$

where α is the learning rate, $\mathcal{L}_{\mathcal{S}_T}$ is the loss on the support set. ❷ *Meta-level Learning*: After task-level learning, the objective is to update the model's parameters θ' such that the model performs well on the query set $\mathcal{Q}_T$:

$$\min_{\theta'} \mathcal{L}_{\mathcal{Q}_T}(f_{\theta'}), \tag{2}$$

where $\mathcal{L}_{\mathcal{Q}_T}$ is the loss on the query set. By iterating over randomly sampled tasks in $\mathcal{T}$, the model f_θ learns to quickly adapt to new sentences (tasks) using only a few IMU signals.

CTC loss for $\mathcal{L}$: Unlike traditional meta-learning where cross-entropy loss is widely used, sign language recognition aligns one sequence (i.e., IMU signals) with another (i.e., a sequence of glosses). Yet, as mentioned earlier, the acquisition of exact alignment is difficult. Therefore, we employ the Connectionist Temporal Classification (CTC) loss [20] to automatically learn the alignment.



(a) High-level illustration of SLR in context of Supervised- and Meta-learning. Colors represent different sentences (tasks) and shapes denote variations due to factors such as signing styles.

(b) An overview of the proposed meta-learning approach for sign language recognition: During task-level learning, we draw random sentences from our dataset and train several models tailored to specific sentences (tasks). In contrast, meta-level learning focuses on training a model that covers the entirety of the sign sentence domain.

Fig. 6: (a) High-level illustration of supervised-learning and meta-learning based sign language recognition. (b) Overflow of *ASLRing*'s meta-learning framework where the ultimate goal is to learn a model that comprehends the entire scope of the sign sentence (task) space.

Given an input IMU signal x (for simplicity, we denote x to be x_i^j) of length N and a target gloss sequence y of length G ($G \leq N$), the CTC loss computes the probability $p(y|x)$ by summing over all possible alignments between x and y.

Let π be an alignment, which is a sequence of labels of length G, including the CTC blank symbol. The probability of π given x is:

$$p(\pi|x) = \prod_{t=1}^{N} y_{\pi_t}^t, \tag{3}$$

where $y_{\pi_t}^t$ is the probability of label π_t at time t in the output of a softmax layer. The total probability $p(y|x)$ is the sum of probabilities of all possible alignments π that can be collapsed to y:

$$p(y|x) = \sum_{\pi:B(\pi)=y} p(\pi|x), \tag{4}$$

where $B(\pi)$ is the function that collapses the sequence π by removing repeated labels and the blank symbol. Overall, the CTC loss is now as follows:

$$\mathcal{L}_{\text{CTC}}(x, y) = -\log p(y|x). \tag{5}$$

IMU Encoder: To encode IMU signals, we employ Transformers [49] architecture because of its efficiency in capturing long dependencies. We detail the transformer structure and discuss the hyperparameter setting in Sec. V-B.

C. Refining Output with N-Gram Models

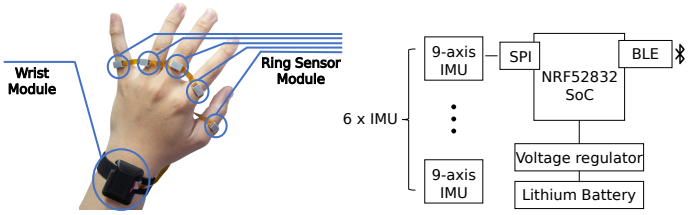
After a model is trained using meta-learning training strategies and during the testing phase, we observed that for a different input signal (e.g., temporal shifted, masked, noise-injected versions), the model still can capture multiple valid glosses of the sentence as shown in Table II. We call this behavior as *partial correctness* of the model because the model has learned certain patterns from the input effectively, but it hasn't completely generalized for the entirety of the input, leading to inconsistent or fragmented outputs on variations. To improve the *partial correctness* behavior, we propose a simple yet effective *output refinement* method based on N-Gram models [12]. As depicted in Algorithm 1, our main idea is to leverage this *partial correctness* behavior: Based on trained model f_θ, we generate a bunch of partially correct sentences, from which we form a bag of glosses $\mathcal{S}$. Since most sentences are partially correct, we argue that the target sentence should share the max number of the glosses with $\mathcal{S}$. To quantify this, we calculate the Intersection over Union (IoU) [42] between a newly generated sentence (from pre-calculated N-Gram model $\mathcal{N}$) and $\mathcal{S}$. Finally, we output the sentence with the max IoU score. Note that we empirically set N to be a combination of several N-gram models where N varies and the rationale is to increase the flexibility in capturing context. We validate our proposed refinement method in Sec. VII.

TABLE II: With varied signal inputs (e.g., temporal shifted, masked, noise-injected versions), the model captures partial glosses of the sentence. However, this *partial robustness* behavior can be improved with the proposed N-Gram models based refinement algorithm (i.e., Algorithm 1). Note that the target for this example is "POSS1 DOG LIKE BALL MOUTH GRAB".

	Prediction	WER (%)
1	GRAB POSS1 DOG	83.3
2	POSS1 MOUTH BALL MOUTH	50.0
3	DOG MOUTH DOG MOUTH	66.6
4	LIKE THROW BALL	75.0
Refinement	DOG LIKE BALL MOUTH GRAB	16.6

Algorithm 1 N-Gram Refinement from Partial Outputs

Input: A set of seen sequences of glosses $\mathcal{G}$, Unseen input x, Model f_θ
Build a Dynamic N-Gram model $\mathcal{N}$ from $\mathcal{G}$.
▷ We set N = {1,2,3} empirically
$\mathcal{V}$ = Variations of x ▷ e.g., masked
for each v in $\mathcal{V}$ **do**
  $\mathcal{S}_v = f_\theta(v)$ ▷ Output glosses
end for
$\mathcal{S}$ = Union of all $\mathcal{S}_v$ ▷ A bag of glosses
for each word w in $\mathcal{S}$ **do**
  Generate a sentence g using $\mathcal{N}$ s.t. length is bounded by $|\mathcal{S}|$
  Add g to $\mathcal{G}_S$
end for
best_score = -1
for each sentence g in $\mathcal{G}_S$ **do**
  score = IoU(g, $\mathcal{S}$) ▷ Overlap ratio as score
  if score > best_score **then**
    best_sentence = g
    best_score = score
  end if
end for
Output: best_sentence



(a) Overview of our sensing platform **(b)** Communication Architecture

Fig. 7: *ASLRing*'s sensing platform in forms of IMU rings on fingers. BLE is used for communication purposes.

V. Implementation

A. ASLRing *Sensing Platform*

The Choice of Sensing Modality: Our goal is to employ a discreet, easily portable, low-cost cost and energy-efficient sensor device, capable of tracking patterns of natural movements of fingers without restriction while being worn throughout the day and night. Encouraged by recent promising work [35], [58], we exploited the Inertial Measurement Unit (IMU) sensors as the sensing method. In comparison to depth cameras [2], [4], IMU sensors are much more robust to lighting, background, and resolution. In addition, IMU sensors have lower power consumption, compact and lightweight form factor, and lower cost.

The Choice of Form Factor: Fig. 7a shows the sensing platform we used in *ASLRing*. We aim to design an unobtrusive, portable, and comfortable devices. Prior works such as sensor gloves [3], impede natural finger motion. EMG armbands [36] require substantial calibration and training datasets. Single ring platform is convenient but is limited in capturing comprehensive data from all fingers [59]. In contrast, *ASLRing* borrows from [60] for its flexibility. The device is designed to emphasize continuous recognition of ASL in entirely natural, unconstrained, and arbitrary signing expressions. Fig. 7b depicts the architecture of the platform. The device consists of five IMU sensors attached to each finger as rings, along with an extra IMU sensor located on the wrist within a smartwatch form factor. The sensing platform's sampling is at 100 Hz. The weight of the platform is 21.2g, contributing to a lightweight form factor and offering comfort to users. A 3.7V, 500mAh LiPo battery is used to supply power with power consumption of 32mA for data streaming, offering battery life about 16 hours.

B. ASLRing *Software Implementation*

ASLRing's software is implemented on a combination of desktop and smartphone devices. Our ML models are implemented with Pytorch [40] library and the training is performed on a desktop with Intel i7-8700K CPU, 16GB RAM memory, and an NVIDIA Quadro RTX 8000 GPU. We use the Adam optimizer [31] with the L2 regularization [11] with a parameter of 0.05 to avoid overfitting issues that may happen in the training process. The learning rate for task-level and meta-level learning is set to 0.007 and 0.003 respectively. Our IMU encoder is based on widely-used transformers [49], and we set the number of attention heads, the number of layers, and the feed-forward dimension to 16, 3, and 256 respectively. We also add dropout [50] with a parameter of 0.4. For the proposed N-gram model based refinement, we set N = {1, 2, 3} considering the tradeoff between accuracy and computation. Note that we opt for the parameters based on simple grid search methods. Once a model is generated from training, the inference is done on a smartphone device using Pytorch Lite [17] on Samsung S20 and OnePlus 9 Pro smartphones.

VI. User Study and Dataset

A. Dataset Preparation

At the beginning of the project, we recruited two native ASL users as interns to create a database with 1080 sentences (English and corresponding glosses) that cover topics such as sports, personal care, arts, and food. For each sentence, a sign video was made by one of the interns for the incoming user study as depicted next.

B. User Study

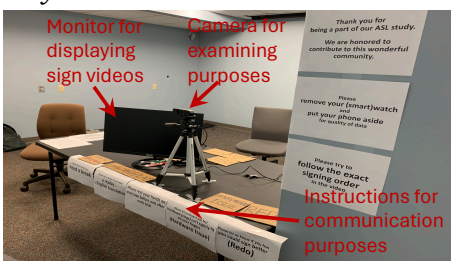


Fig. 8: Example of user study setups: users were seated in front of the monitor and our team members were seated in the opposite. Since we are not fluent in ASL, we prepared instructions and cardboard for communication purposes.

In our user study, 14 Deaf users (9 females, 5 males) with native ASL fluency were recruited. The users are

aged between 20-50 and weigh from 50 to 90 kg and the IRB committee has approved the study. Note that because of the difficulty of recruiting native ASL users, our user study is not in one go. Instead, we schedule the study multiple times for different users, thus the actual user study is across three months. Fig 8 depicted our user study setup in a conference room.

The study procedure is as follows. Pre-made sign videos will be played and users need to understand and re-sign the same content from the sign videos while they wear *ASLRing*'s sensing device on both hands with the sensors snugly fit on fingers as depicted in Fig. 7a. Each user may take 3 or more breaks (per user's desire) and each break is at least 3 minutes. The sensing device will be removed and remounted for breaks and the total study time for each user is two hours. Due to different cultural backgrounds, communities, etc., we do not ask users to follow the exact signs as long as similar content is guaranteed. Although this demonstrates our respect for the ASL community and allows users to freely express themselves, it also brings challenges (e.g., complex articulation, and signing styles) as in Sec. III.

C. Dataset Summary

Unlike prior works [33], [56] that simplify glosses or capture only parts of finger motions, we keep the original glosses created by the ASL interns and capture all finger motions for comprehensiveness. After simple data cleaning to remove bad data due to hardware issues, a total number of 4828 sentences from 14 users were acquired. Among these data, we have 1018 unique sentences with a maximal length of 16 glosses of daily conversation that cover topics in sports, personal care, arts, food, etc., and the vocabulary size is 1057. On average, each unique sentence is signed by 5-6 users. And each user signs around 200 distinct sentences in two hours. Table III depicts a comparison between *ASLRing* and prior works. Evidently, our dataset is comprehensive and more realistic with native ASL users.

TABLE III: *ASLRing* is more comprehensive than prior systems.

System	No. of Native ASL Users (Test)	Vocabulary Size	Distinct Sentence Signed Per User
DeepASL [18]	0	56	100
MyoSign [55]	0	16	48
SignSpeaker [23]	0	103	73
FinGTraC [35]	0	90	50
DeepSLR [52]	0	51	60
SonicASL [28]	0	42	30
WearSign [56]	0	100	16
SignRing [33]	2	934	15
ASLRing	14	1057	200

VII. Performance Evaluation

We present the evaluation results for *ASLRing* and the organization is as follows. ■ We first elaborate on training and testing data. ■ Then, we briefly introduce the evaluation metric that is widely used in the sign language recognition task. ■ Overall performance (e.g., on individual users) of *ASLRing* is presented. ■ A robustness study is conducted to verify that *ASLRing* works across various situations such as users' age, various signing speeds, and different environments. ■ An ablation study is conducted to verify the design choice of *ASLRing*. ■ A challenging case (unseen sentences) is studied along with our insights. ■ We present qualitative results to demonstrate *ASLRing*'s performance. ■ And lastly, we evaluate *ASLRing* on smartphones for the power consumption details.

A. Training Data and Testing Data

Training Data: As described in Sec. IV, we employed meta-learning as our main framework, in which we defined each unique sentence as a task. Therefore, during the training process, we randomly sample tasks (sentences) at each iteration, and for each task, we further divide IMU signals into support and query sets for task-level and meta-level learning respectively. We randomly sample 10% and 70% in a task for support and query set, and the rest 20% for testing purposes.

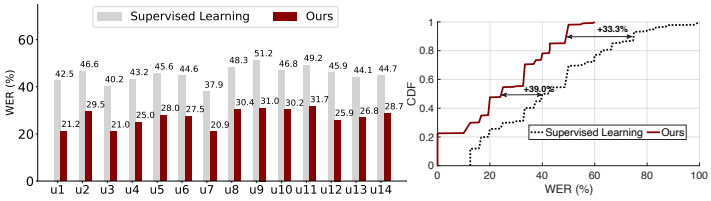
Testing Data: After the model is trained on different tasks, we acquired the model that can generalize well on the same task with various inputs or different tasks. Correspondingly, we have two types of testing data: ❶ Unseen users: sentences from unseen users are the sentences that are signed by a different user who is not seen in the training. ❷ Unseen sentences: the sentences are not available in the training process. Note that our results presented next are based on unseen users, and we present the results for unseen sentences separately.

B. Metrics

To evaluate *ASLRing*'s performance, like the prior works [33], [56], we employ word error rate (WER). WER is a standard metric utilized in the field of speech and language processing to assess the performance of automatic speech recognition (ASR) systems. The formula for WER is derived by considering the number of substitutions (S), insertions (I), and deletions (D) required to match the system's output to a reference transcript. It's computed relative to the number of words in the reference (N): WER $= \frac{S+I+D}{N}$ The metric offers a holistic measure of a system's accuracy, as it takes into account all possible discrepancies between the predicted output and the actual reference. Note that lower WER indicates better performance of the system.

C. Overall Performance

As depicted in Sec. III, we observed several challenging cases. One of them is user diversity, which includes diversities from different cultural backgrounds, signing habits, etc. This imposes difficulties in creating a generalized model well on all the other unseen users as also reported by prior works [27], [30]. Therefore, to incorporate these diversities, *ASLRing*, similar to other systems, also needs an additional fine-tuning process for unseen users (after the leave-one-user-out training process). The amount of data needed for fine-tuning is set to 15% ($\approx$ 30 sentences)



(a) Performance on unseen users **(b)** WER distribution

Fig. 9: (a) *ASLRing* performs over 40.1% well on unseen users compared to supervised learning with the same amount of fine-tuning data. (b) WER distribution for *ASLRing* and supervised learning based SLR system. 39.0% and 33.3% improvement are achieved for the median and 90%-ile accuracy, respectively.

for both *ASLRing* and its supervised learning counterpart. And we give more details in Sec.VII-E. Note that for comparison purposes, we build a supervised learning counterpart with the same architectures (as *ASLRing*) but different training strategies.

As shown in Fig. 9a, *ASLRing* performs better than its supervised learning counterpart with a 40.1% boost on average. The accuracy variation of the individual user could happen because of the various cultural backgrounds. Our pre-made sign videos are recorded by our ASL intern who is a black woman, while some users (e.g., u9) from the user study have some Asian backgrounds. And since we are not restricting the users to sign exact signs in our user study as mentioned in Sec. VI, users tend to sign in their own styles, resulting in various individual performances even though the meta-learning training scheme with careful defined tasks was designed to decouple variation from signals, user styles are worthy to dig deeper. Nevertheless, we believe *ASLRing* is stable across unseen users with an average 26.9% WER.

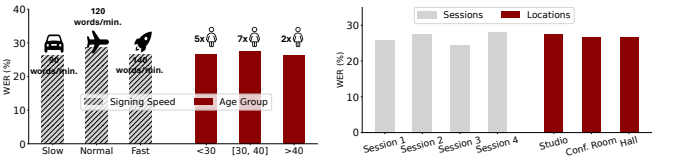
Fig. 9b, on the other hand, depicted the distribution of WER on the sentences signed by unseen users. Compared with its supervised learning counterpart, *ASLRing* achieved 39.0% and 33.3% improvement for the median and 90%-ile WER respectively, demonstrating the effectiveness of meta-learning based sign language recognition system. While the overall performance is reasonable, we found the span of the performance is large, e.g., WER ranges from 0% to 60%. We believe one reason is replaceable signs or various signing orders that happen in our user study. This results in the inconsistency between our labels (although some of them have similar semantic meanings, the CTC loss cannot accommodate this inconsistency) and IMU signals (e.g., one example is shown in Sec. III), which could be one factor to mislead the learning process. We leave this deep clean or more advanced signal processing for future work as discussed in Sec. VIII.

D. Robustness Study

Study for Various Speed: To validate the impact of signing speeds on sign language recognition accuracy, we categorize users into three classes (i.e., ≈90 words/minute, ≈120 words/minute, and ≈140 words/minute.) based on sentence lengths (i.e., the number of glosses in a sentence) and the corresponding video lengths. Evidently, as shown in Fig. 10a, *ASLRing* can adapt to different signing speeds. We believe this is due to the transformer architecture we employed, in which the temporal correlation is learned effectively.

Study for Age Group: In order to study the recognition accuracy of different age groups, we roughly split our users (aged from 20 to 50) into three groups, representing potential variations from living environments. As shown in Fig. 10a, *ASLRing* is robust to these variations. Thanks to the well-considered task definition, we believe the employed meta-learning based framework can handle such variations from signals.

Robustness to Sensor Positions and Orientations: In the study, users can have more than 3 breaks. For these breaks, we remove and remount the sensing device from and onto users to validate *ASLRing*'s robustness to varied sensor positions/orientations with respect to the human body. Fig. 10b depicts the accuracy across sessions (breaks). Evidently, the recognition accuracy doesn't get affected and is stable across 4 sessions (3 breaks). We believe our sensing device can fit snugly onto users' fingers. Therefore, any minor variation in positions across breaks is small, leaving a negligible impact on accuracy.

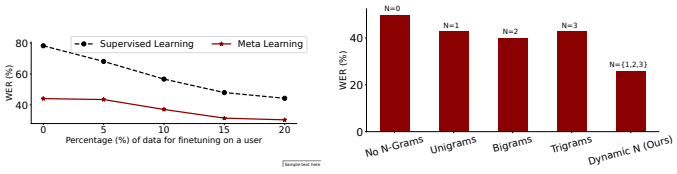


(a) Signing Speed and Age Group **(b)** Sessions and Locations

Fig. 10: (a) *ASLRing* is adapt to various signing speeds and age groups (b) *ASLRing* is robust to various sensor locations when removing and remounting devices during breaks in the study and is ubiquitous to the environments that have potential magnetic inferences from refrigerators, cameras, projectors, etc.

Robustness to Environment: Note that the user study was set up in different environments and each place has different surroundings (refrigerators, lights, cameras, projectors, etc.), potentially affecting the orientation estimation for our IMU sensors due to magnetic inferences. We validate *ASLRing*'s performance on this variation in Fig. 10b. Evidently, *ASLRing* is ubiquitous to any environment due to the opportunistic calibration technique from A3 [61], where the gravity vector and magnetic north vector are opportunistically determined and used to reset the drift error in single-integral based gyroscope orientation estimation.

E. Ablation Study

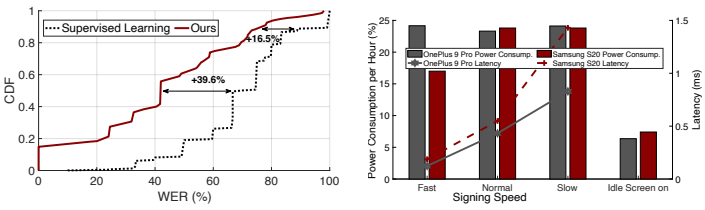


(a) percentage of data used to fine-tune under different training strategies **(b)** Recognition accuracy on varied N-Gram models

Fig. 11: (a) Meta learning based training strategies outperform that of supervised learning by a large margin when a varied number of fine-tuning data is available. (b) Our proposed dynamic N-Gram refinement method is accurate than any of the other models because of increased flexibility on choosing next word.

Comparison Between Meta Learning and Supervised Learning employed in Prior Works: Fig. 11a depicts the recognition accuracy as a function of the size of fine-tuning data from an unseen user. When there is no data (i.e., 0%) from given unseen users, we note that meta-learning based model still outperforms its supervised learning counterpart, which has been utilized in prior works [18] [55] [23] [35]. We conclude the reason is that with supervised learning, the performance is greatly dependent on the quality and the diversity of the training data, so if there is one unseen user whose data are not drawing from the same distribution of the training data, supervised learning based systems usually suffer from generalization problems. In contrast, thanks to the definition of tasks in meta-learning, *ASLRing* focuses more on the variation of input signals for the same sentence and the generalization on a wide span of sentences (tasks), thus leading to a sweet spot that works for variations and all tasks (sentences). We believe this is also the reason why it can quickly learn with a small amount of fine-tuning data, achieve convergence, and even outperform supervised learning when only a small number of fine-tuning data is available. Based on this result, we use 15% data for *ASLRing* adapting to new users. Note that 15% of data corresponds to 30 sentences and it roughly takes 18 minutes to collect according to our user study. Yet, we believe the amount of time needed for collecting these 15% data can be much shorter on the user's end (without the complex collection procedure in our study). Even though more fine-tuning data might lead to a more personalized model, we leave this opportunity for future work. *ASLRing* utilizes meta-learning training strategies and achieves better recognition accuracy compared to its supervised learning counterpart no matter whether there is no or a small number of data to finetune.

Refining Output with N-Gram Models: Our proposed refinement is based on observation during testing: when input signals get disturbed (e.g., temporal shifted, masked), the trained model tends to output partially correct glosses. To validate the effectiveness of the proposed refinement method, we conducted an experiment where we tested N-Gram models when N varied from 0 to 3 and compared the results with our proposed dynamic N-Gram refinement method. As depicted in Fig. 11b, our proposed method outperforms other basic models by a large margin. This is because by combining different N-gram models (i.e., N={1, 2, 3}), we could capture more context, thus increasing the flexibility (more options) to choose the next most reasonable word. Specifically, lower-order N-grams (like unigrams and bigrams) are more general and have broader coverage, while higher-order N-grams can capture more specific and nuanced patterns in the glosses. Combining them ensures that both general patterns and specific nuances are taken into account. However, we stopped at N=3 due to computation efficiency, and left incorporating large language models for refinement to future work as discussed in Sec. VIII.



(a) Unseen Sentences **(b)** Power Consumption

Fig. 12: (a) *ASLRing*'s performance on unseen sentences: A median ≈40% WER is achieved, which is a 39.6% improvement compared to that of supervised learning. More details are in the text. (b) Power consumption of *ASLRing*.

F. Challenge Case: Performance on Unseen Sentence

Unseen sentences have been a challenging task in sign language recognition. Prior works [56] [33] tackled this task by increasing the diversity of datasets or by synthesizing unseen sentences from individual words. Yet, increasing diversity via data collecting is expensive, and harvesting virtual data from other sources (e.g., sign videos) inherits existing issues from the videos (e.g., occlusion, low resolution, overexposure), hindering the performance of downstream tasks. *ASLRing* takes a different branch by employing meta-learning training strategies, in which each sentence is viewed as a task. After being trained, the model acquires a comprehensive semantic knowledge spanning the task spaces. As depicted in Fig. 12a, without fine-tuning, *ASLRing* achieves a median WER ≈40%, which is improved by 39.6% compared to its supervised learning counterpart. We note that the accuracy of *ASLRing* on unseen sentences is much lower than that of *ASLRing* on unseen users. We conclude that the reasons are two-fold: ❶ Articulation: as mentioned in Sec. III, native ASL users have fluent and complex articulation patterns and varied styles. This makes learning the segmentation of individual words hard even though we employed widely-used CTC loss to automatically learn the boundaries. Without boundaries between words, a new sentence is hard to predict. ❷ Single Sentence as Task: when adopting meta learning training strategies, *ASLRing* treats each sentence as a task. While this has demonstrated a way to decouple variations (e.g., from signing styles, and environmental factors) from actual signals to achieve a "neutral style", the prediction for an unseen sentence that contains glosses from multiple other seen sentences (in training), tends to lean toward one of the seen sentences, resulting in higher word error rate because the high-level semantic meaning of glosses is not well-integrated into the learning process. Thus, considering the semantic meaning of sentences when splitting them into tasks is a potential solution to improve the performance of unseen sentences. We leave this for future work as discussed in Sec. VIII. Nevertheless, we believe *ASLRing* demonstrates a reasonable performance on unseen data with native ASL users.

G. Power Consumption

The power consumption of the sensor device itself is discussed in Sec. V-A. Here, we analyze the power consumption of executing ML models in *ASLRing* by profiling the power consumption of the models with Batterystats and Battery Historian [7]. As depicted in Fig. 12b, the average real-time continuous power discharge rate is 21.5% and 23.8% per hour for Samsung S20 and OnePlus 9 Pro. In real world, the power discharge rate should be much lower as users do not continuously sign. The average latency of execution of ML models on the two mobile devices utilizing CPU is approximately 722 ms and 461 ms respectively. As average lag times of real sign language interpreter are 2-3 seconds [10], *ASLRing* should be sufficient for real-time applications. For reference, the idle display-screen on discharge rate is 7.41% and 6.35% per hour for Samsung S20 and OnePlus 9 Pro. We plan to optimize our ML models with deep compression [21] to reduce power consumption.

H. Qualitative Results

Table IV depicts partial recognition results of the unseen user and unseen sentence. Evidently, *ASLRing* recognizes signs and maps them into corresponding glosses accurately because *ASLRing* employs meta learning based training strategies to handle various input signals of the same sentences and proposes an N-Gram models based refinement method to improve the prediction results. Overall, we believe the results are encouraging.

VIII. Discussion and Future Work

Pitfalls of Single-sentence Task of Meta-Learning and Potential solution: *ASLRing* utilized meta learning strategies to enhance the recognition results from various user inputs. Although reasonable performance is demonstrated, we found several challenging cases. Firstly, as discussed in Sec. III, native ASL users have their own signing styles and understanding of a sentence. For instance, users can replace signs with a different sign as long as the meaning is similar (e.g., sign for "LOVE" and "FALL-IN-LOVE"). This results in the inconsistency between the labels (glosses) and IMU signals that mislead the learning process when recognition systems are greatly based on the data. Secondly, *ASLRing* struggles with unseen sentences and the reasons are 1) unclear boundaries due to highly articulated signs by native users; 2) due to only a few words overlapping between unseen sentences and some of the trained sentences, the prediction of unseen sentences will base on one of the trained sentences, leading to a higher word error rate. Based on these challenges, we realized that when defining tasks, we simply treat each sentence as a task individually without considering the external knowledge of the semantic meanings of the sentences. Thus, to improve *ASLRing* on these challenging tasks, we could define a task that has cluster sentences that share similar meanings. As depicted in Fig. 13, sentences with similar contents can be grouped in the embedding space, suggesting that *ASLRing* can leverage external semantic knowledge from language models to build a semantic-aware meta-learning network to address the above-mentioned challenges.

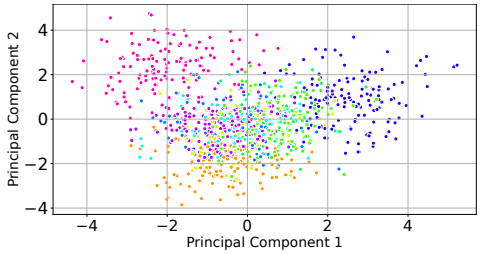


Fig. 13: We embed 1080 sign sentences using publicly available language models, group the embeddings into several clusters via K-Means [22] and project these clusters onto a 2D plane via PCA [53]. These clusters suggest external semantic knowledge from language models can be utilized to address challenges such as unseen sentences, and inconsistency between signs and labels.

Utilizing foundation models to enhance prediction accuracy: While *ASLRing* has demonstrated improvement through the adoption of a simple N-gram-based model refinement approach, it primarily relies on insights from our self-collected dataset. In the expansive realm of natural language processing, foundation models (like Large Language Models or LLMs) have come to the forefront. Owing to the prowess of LLMs in comprehending sentence contexts, emotions, and intricate linguistic subtleties, integrating foundation models can lead to predictions that are not only more precise but also attuned to context, thanks to their vast corpus. Even though *ASLRing* harnesses meta-learning to interpret results from previously unseen data, merging it with foundation models might offer superior accuracy. This is due to the enhanced contextual understanding and grasp of linguistic nuances provided by the foundation models. We seek this opportunity in the future.

IX. Related Work

A. Sign Language Recognition

Visions: Recent works [34], [57] use RGB cameras to recognize sign language. SignBERT [24] and HMA [25] propose to decode 3D hand key points of sign language from RGB videos. Ye et al. [54] propose a hybrid 3D recurrent convolutional neural network (3DRCNN) to recognize ASL gestures from continuous videos via a Kinect sensor. Camgoz et al. [16] propose a sequence-to-sequence learning methodology using a series of specialized expert systems for sign language recognition. However, vision-based systems have a critical weakness in terms of raising privacy concerns and conditional light environment [15]. In contrast, *ASLRing*'s solution is resilient to environmental variables such as occlusion and lighting, as well as minimizing privacy concerns.

Wearables: Savur et al. [45] propose an ASL recognition system using the eight-channel surface Electrocardiography (sEMG) to recognize alphabet letters, words, and sentences. However, these recent studies focus on wearing

TABLE IV: Qualitative recognition results of *ASLRing*. Note that red denotes missing words from targets and blue denotes extra words from targets. While some glosses are hard to catch due to factors such as co-articulated signs, we believe the overall recognition result is encouraging.

	Prediction	Target	WER (%)
Unseen Users	COOKIE TYPE IX FAVORITE WHICH	COOKIE TYPE IX FAVORITE WHICH	0.0
	COLD SOUP MATCH PERFECT	COLD SOUP MATCH PERFECT	0.0
	CURIOUS LEARN JAPANrep GESTURE	CURIOUS LEARN JAPANrep GESTURE QQ	20.0
	DRINKalc DELICIOUS	BAR CHARACTERISTIC DRINKalc FOOD DELICIOUS	60.0
Unseen Sentences	SODA PRICE HOW-MUCH	SODA PRICE HOW-MUCH	0.0
	READ BOOK LIKE	READ BOOK LIKE	0.0
	TOPUsym FAVORITE COMMUNITY WHICH	POSS2 TOPUsym FAVORITE COMMUNITY WHICH	20.0
	EAT NOON POSS DAD	BING SALAD EAT NOON WORK ALWAYS	66.6

gloves [6], [32], which hinders users from performing natural and dexterous activities with fine precision as investigated in recent work [43]. SignSpeaker [23] recognizes 73 sentences constructed by 103 distinct words via a smartwatch. WearSign [56] constructs an ASL dataset of 250 sentences from 15 student volunteers. FinGTrAC [35] shows the feasibility of recognizing the 100 words via a smart ring and a smartwatch. SignRing [33] synthesizes virtual IMU data based on a two-view triangulation tracking approach. Yet the method still suffers from inherent drawbacks in videos such as occlusions, and overexposure. While these systems made efforts for sign language recognition in terms of increasing data diversity, some simplifications have been proposed such as reducing the complexity of sign sentences [33], and only capturing signs from one hand (both hands are important in sign languages) [33], [56], hindering the wide adoption. When validating, most of them can only recruit ASL learners. In contrast, *ASLRing* employs a low-cost, lightweight, and wireless sensing platform to push the boundary of wearable-based solutions with a database with 1080 sentences constructed from a vocabulary size of 1057 from 14 native ASL users.

Radio Frequency (RF) Signals: Recent works combine wireless channel state information (CSI) and Doppler shifts to track the motion of the hand and classify discrete gestures [46], [48]. [51] proposes a multi-view deep neural network, fusing micro-doppler in different directions, for Chinese sign language (CSL) recognition. SignFi [37] recognizes sign language using CSI from WiFi APs. ExASL [44] tracks point clouds computed from the range-doppler spectrum and angle of arrival spectrum of mmWave radars, showing the feasibility of classifying up to 23 discrete ASL gestures via mmWave. In contrast to prior works only performing predefined sign recognition, *ASLRing* recognizes continuous sign language by 14 native users with an active and more pervasive sensing platform that closely captures signs without range limitation.

B. Meta learning

The core idea behind meta-learning is to design algorithms that can rapidly adapt to new tasks with minimal data, leveraging prior knowledge acquired from related tasks. Finn et al. [8] presented Model-Agnostic Meta-Learning (MAML), a seminal approach that's designed to be applicable across various models and tasks. MAML initializes a model in such a way that a few gradient steps on a new task will lead to effective adaptation. While meta-learning primarily focuses on rapid adaptation, its principles share common ground with transfer and multi-task learning. Caruana [14] previously discussed the value of multi-task learning in improving the generalization of models by sharing representations across tasks. In the span of sign language recognition, [26] proposes a contrastive disentangled meta-learning frame to decouple variations from users to achieve a signer-independent sign language translation model. Similarly, [27] leverages meta-learning to minimize the data for adapting the model to new users. *ASLRing* embodies the essence of meta-learning for sign language recognition. Learning from interactions with native ASL users, *ASLRing* distinctly treats sign sentences as distinct tasks, allowing it to separate variations from input signals. Consequently, with the acquired knowledge spanning various sentence spaces, *ASLRing* is equipped to predict previously unseen data.

X. Conclusion

We propose *ASLRing*, a meta-learning-based recognition system that addresses the existing diversity issues in sign languages. By involving meta learning in SLR, we observe encouraging results when facing challenge tasks such as unseen users and unseen sentences. We also discussed future steps for improving *ASLRing* on challenging cases by leveraging semantic knowledge of signs from some external large language models. In contrast to prior works, *ASLRing* is evaluated with 14 native ASL users. *ASLRing* achieves 26.9% word error rate (WER) across users, and a ≈40% WER on unseen sentences. Furthermore, the accuracy is also consistent across variations in signing speed, sensor wearing positions, and dialects/accents, demonstrating the robustness of *ASLRing*.

Acknowledgements

Our gratitude goes to the anonymous reviewers for their invaluable feedback. This research has been partly funded by NSF grants: CAREER-2046972.

REFERENCES

[1] How many deaf people are there in united states. https://research.gallaudet.edu/Demographics/deaf-US.php.

[2] Leap motion. https://developer.leapmotion.com/, 2012.

[3] Cyberglove. http://www.cyberglovesystems.com/, 2017.

[4] Kinect2.0. https://developer.microsoft.com/en-us/windows/kinect, 2021.

[5] World health organization. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss, 2021.

[6] AHMED, M. A., ET AL. A review on systems-based sensory gloves for sign language recognition state of the art between 2007 and 2017. *Sensors* (2018).

[7] Profile battery usage with batterystats and battery historian. https://developer.android.com/topic/performance/power/setup-battery-historian, 2021.

[8] ANTONIOU, A., EDWARDS, H., AND STORKEY, A. How to train your maml. *arXiv preprint arXiv:1810.09502* (2018).

[9] BAHAN, B. J. Non-manual realization of agreement in american sign language.

[10] BARIK, H. Simultaneous interprétâtion: Temporal & quantitative data. ll thurstone psychometric laboratory. Tech. rep., 1972.

[11] BERTERO, M., ET AL. The stability of inverse problems. In *Inverse scattering problems in optics*. Springer, 1980.

[12] BROWN, P. F., ET AL. Class-based n-gram models of natural language. *Computational linguistics* (1992).

[13] CAI, Y., ET AL. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *ECCV* (2018).

[14] CARUANA, R. Multitask learning. *Machine learning 28* (1997), 41–75.

[15] CHEN, A. T.-Y., ET AL. Context is king: Privacy perceptions of camera-based surveillance. In *IEEE AVSS* (2018).

[16] CIHAN CAMGOZ, N., ET AL. Subunets: End-to-end hand shape and continuous sign language recognition. In *ICCV* (2017).

[17] FACEBOOK. Introduce lite interpreter workflow in Android and iOS. "https://pytorch.org/mobile/android/", 2021.

[18] FANG, B., ET AL. Deepasl: Enabling ubiquitous and non-intrusive word and sentence-level sign language translation. In *ACM Sensys* (2017).

[19] GOWDA, M., ET AL. Bringing iot to sports analytics. In *NSDI 17* (2017).

[20] GRAVES, A., ET AL. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ACM ICML* (2006).

[21] HAN, S., MAO, H., AND DALLY, W. J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149* (2015).

[22] HARTIGAN, J. A., AND WONG, M. A. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics) 28*, 1 (1979), 100–108.

[23] HOU, J., ET AL. Signspeaker: A real-time, high-precision smartwatch-based sign language translator. In *MobiCom* (2019).

[24] HU, H., ET AL. Signbert: pre-training of hand-model-aware representation for sign language recognition. In *ICCV* (2021).

[25] HU, H., ZHOU, W., AND LI, H. Hand-model-aware sign language recognition. In *AAAI* (2021).

[26] JIN, T., ET AL. Contrastive disentangled meta-learning for signer-independent sign language translation. In *ACM MM* (2021).

[27] JIN, T., ZHAO, Z., ZHANG, M., AND ZENG, X. Mc-slt: Towards low-resource signer-adaptive sign language translation. In *ACM MM* (2022).

[28] JIN, Y., ET AL. Sonicasl: An acoustic-based sign language gesture recognizer using earphones. *ACM IMWUT* (2021).

[29] JIN, Y., ET AL. Smartasl: " point-of-care" comprehensive asl interpreter using wearables. *ACM IMWUT* (2023).

[30] JIN, Y., ET AL. Transasl: A smart glass based comprehensive asl recognizer in daily life. In *IUI* (2023).

[31] KINGMA, D. P., ET AL. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[32] KURODA, T., ET AL. Consumer price data-glove for sign language recognition. In *Proc. of 5th Intl Conf. Disability, Virtual Reality Assoc. Tech., Oxford, UK* (2004).

[33] LI, J., ET AL. Signring: Continuous american sign language recognition using imu rings and virtual imu data. *ACM IMWUT* (2023).

[34] LIKHAR, P., BHAGAT, N. K., AND RATHNA, G. Deep learning methods for indian sign language recognition. In *IEEE ICCE-Berlin* (2020).

[35] LIU, Y., ET AL. Finger gesture tracking for interactive applications: A pilot study with sign languages. *ACM IMWUT* (2020).

[36] LIU, Y., ET AL. Wr-hand: Wearable armband can track user's hand. *ACM IMWUT 5*, 3 (2021), 1–27.

[37] MA, Y., ET AL. Signfi: Sign language recognition using wifi. *ACM IMWUT* (2018).

[38] MUELLER, F., ET AL. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *IEEE CVPR* (2018).

[39] ORGANIZATION, W. H. Deafness and hearing loss. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss.

[40] PASZKE, A., ET AL. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS* (2019).

[41] PICHLER, D. C. Word order variation and acquisition in american sign language.

[42] REZATOFIGHI, H., TSOI, N., GWAK, J., SADEGHIAN, A., REID, I., AND SAVARESE, S. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR* (2019).

[43] RODA-SALES, A., ET AL. Effect on manual skills of wearing instrumented gloves during manipulation. *Journal of biomechanics* (2020).

[44] SANTHALINGAM, P. S., ET AL. Expressive asl recognition using millimeter-wave wireless signals. In *IEEE SECON* (2020).

[45] SAVUR, C., AND SAHIN, F. American sign language recognition system by using surface emg signal. In *IEEE SMC* (2016).

[46] SHANG, J., AND WU, J. A robust sign language recognition system with multiple wi-fi devices. In *MobiArch Workshop* (2017).

[47] SIM, D., ET AL. Low-latency haptic open glove for immersive virtual reality interaction. *Sensors* (2021).

[48] TAN, S., AND YANG, J. Wifinger: Leveraging commodity wifi for fine-grained finger gesture recognition. In *ACM Mobihoc* (2016).

[49] VASWANI, A., ET AL. Attention is all you need. *NeurIPS* (2017).

[50] WAGER, S., ET AL. Dropout training as adaptive regularization. In *NeurIPS* (2013).

[51] WANG, X., ET AL. Chinese sign language recognition based on multiview deep neural network for millimeter wave radar. In *Artificial Intelligence and Machine Learning in Defense Applications IV* (2022).

[52] WANG, Z., ET AL. Hear sign language: A real-time end-to-end sign language recognition system. *IEEE Transactions on Mobile Computing* (2020).

[53] WOLD, S., ESBENSEN, K., AND GELADI, P. Principal component analysis. *Chemometrics and intelligent laboratory systems* (1987).

[54] YE, Y., ET AL. Recognizing american sign language gestures from within continuous videos. In *IEEE CVPR Workshops* (2018).

[55] ZHANG, Q., ET AL. Myosign: enabling end-to-end sign language recognition with wearables. In *IUI* (2019).

[56] ZHANG, Q., ET AL. Wearsign: Pushing the limit of sign language translation using inertial and emg wearables. *ACM IMWUT* (2022).

[57] ZHAO, K., ET AL. Real-time sign language recognition based on video stream. *International Journal of Systems, Control and Communications* (2021).

[58] ZHOU, H., ET AL. Learning on the rings: Self-supervised 3d finger motion tracking using wearable sensors. *ACM IMWUT* (2022).

[59] ZHOU, H., ET AL. One ring to rule them all: An open source smartring platform for finger motion analytics and healthcare applications. In *ACM/IEEE IoTDI* (2023).

[60] ZHOU, H., ET AL. Signquery: A natural user interface and search engine for sign languages with wearable sensors. In *ACM MobiCom* (2023).

[61] ZHOU, P., ET AL. Use it free: Instantly knowing your phone attitude. In *ACM MobiCom* (2014).

[62] ZHUANG, W., ET AL. Design of human activity recognition algorithms based on a single wearable imu sensor. *IJSNET* (2019).