

# SignGlass: First-Person View Comprehensive and Generalizable ASL Translation Using Wearable Glass

Yongxiang Cai\*

Binghamton University, State  
University of New York  
Binghamton, USA  
ycai3@binghamton.edu

Taiting Lu\*

Pennsylvania State University  
University Park, USA  
txl5518@psu.edu

Zhenghao Li

Pennsylvania State University  
University Park, USA  
zkl5382@psu.edu

Hao Zhou

Pennsylvania State University  
University Park, USA  
hao.zhou@psu.edu

Kenneth DeHaan

Gallaudet University  
District of Columbia, USA  
kenneth.dehaan@gallaudet.edu

Xuhai Xu

Columbia University  
New York, USA  
xx2489@columbia.edu

Mahanth Gowda

Pennsylvania State University  
University Park, USA  
mahanth.gowda@psu.edu

Yincheng Jin

Binghamton University, State  
University of New York  
Binghamton, USA  
yjin5@binghamton.edu

## ABSTRACT

Communication barriers between Deaf and Hard-of-Hearing (DHH) individuals and hearing individuals remain a major challenge, highlighting the need for technologies that enable seamless sign language interpretation. However, current American Sign Language (ASL) recognition and translation systems face key limitations, including poor portability, complex usage settings, incomplete capture of essential components, and weak generalization, reducing practicality and user acceptance.

To address these challenges, we present SignGlass, the first smart glasses equipped with three wearable cameras for comprehensive capture of both manual and non-manual ASL markers, supported by advanced algorithms for real-time recognition and translation into English. Specifically, SignGlass integrates a jitter-aware spatio-temporal attention mechanism for robust recognition of hand movement patterns. Complementing this, a dual-camera, temporally-aware facial module captures the subtle facial expressions essential for ASL comprehension. To support diverse signing styles across individuals, we further introduce a cascaded data augmentation strategy to improve model generalization.

In a user study with 14 Deaf participants, SignGlass achieved high translation accuracy and was well-received, demonstrating its effectiveness in bridging communication gaps. This work highlights

the promise of multi-camera wearable systems in advancing ASL translation and promoting more accessible communication.

## CCS CONCEPTS

• **Human-centered computing** → **Interaction techniques**.

## KEYWORDS

ASL Recognition and Translation; Smart Glass; Wearable Cameras

### ACM Reference Format:

Yongxiang Cai, Taiting Lu, Zhenghao Li, Hao Zhou, Kenneth DeHaan, Xuhai Xu, Mahanth Gowda, and Yincheng Jin. 2025. SignGlass: First-Person View Comprehensive and Generalizable ASL Translation Using Wearable Glass. In *The 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*, September 28–October 1, 2025, BUSAN, KOREA. ACM, New York, NY, USA, 17 pages. <https://doi.org/10.1145/xxxx>

Open-source SignGlass Project Repository.

## 1 INTRODUCTION

It is estimated that over 70 million individuals worldwide are deaf [57, 74]. Within this context, sign languages play a crucial role in communication for Deaf and Hard-of-Hearing (DHH) communities, often serving as their primary language [21]. Among them, American Sign Language (ASL) is the most widely used sign language, particularly in North America [40]. ASL is a structured visual language that comprises both manual and non-manual markers to convey complex linguistic information. Manual markers include four fundamental elements—handshape, palm orientation, movement, and location—which form the core structure of the language. Non-manual markers, such as facial expressions and head or body postures, contribute to grammatical structure and semantic nuance [54, 72]. For example, the manual sign for "I UNDERSTAND" involves specific hand movements, but incorporating raised eyebrows and a slight head tilt—both non-manual markers—transforms the

\*Yongxiang Cai and Taiting Lu contributed equally.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
UIST '25, September 28–October 1, 2025, BUSAN, KOREA

© 2025 Association for Computing Machinery.

ACM ISBN 979-8-xxx-xxxx-x/xx/xx...\$15.00

<https://doi.org/10.1145/xxxx>

meaning into a question: "Do I understand?" Additionally, ASL follows a unique set of linguistic rules that differ significantly from English [41, 76]. Missing any linguistic component—manual, non-manual, or grammatical—can hinder comprehension and lead to communication breakdowns [22]. Therefore, accurate and effective ASL interpretation is essential for facilitating clear communication between DHH individuals and hearing individuals.

Traditionally, professional ASL interpreters have played an indispensable role in enabling communication between DHH individuals and hearing people. However, their high cost and limited availability pose substantial barriers, particularly in spontaneous interactions. This limitation underscores the pressing need for accessible, cost-effective, and user-friendly ASL recognition and translation technology that can provide real-time assistance to DHH individuals in their daily lives. To address this challenge, prior research has explored ASL recognition through two primary technological approaches: camera-based methods and wearable sensor-based systems. Camera-based methods rely on computer vision techniques for sign language recognition using video-based approaches [14, 23, 35, 56, 66]. These systems have achieved high recognition accuracy [18, 24], yet camera-based systems require an unobstructed frontal view of the user's gestures, making them impractical for real-time communication in dynamic or crowded environments, such as public spaces and classrooms. Meanwhile, wearable sensor-based systems incorporate various sensor modalities, such as electromyography (EMG) for detecting muscle activity [71, 76, 77], inertial sensors for motion tracking in smart rings or wristbands [44, 48, 79, 80], and acoustic sensors for sound-based sign detection [37, 40]. However, wearable systems often fail to capture non-manual markers—such as facial expressions, head movements, and body posture—which are essential for accurate ASL interpretation [79, 80]. Thus, recent solutions like SmartASL and TransASL [37, 41] use multiple wearable devices, such as smartwatches and earphones, to recognize both manual and non-manual markers.

However, these systems struggle to be effectively deployed in real-life scenarios due to their limited generalizability, rendering them unreliable for new users, as evidenced by their leave-one-user-out Word Error Rate (WER) exceeding 75% [37, 41]. To achieve feasible performance, a substantial amount of additional personalized data is required to retrain the existing model. However, collecting sufficient personalized ASL data demands tremendous effort, making daily usage impractical. This is largely because in-ear sensing for facial gesture recognition fails to deliver consistent performance across different users [19, 37, 39, 41, 70]. Furthermore, recent wearable systems that employ acoustic sensing [38, 40, 46], EMG sensing [76, 77], and inertial measurement unit (IMU) sensing [44, 46] generate lower-dimensional features compared to image-based methods, limiting their capacity to capture and recognize the complex components of ASL, especially for fine-grained motions. Additionally, the multiple-device settings significantly increase system complexity and reduce usability. The unconventional use of earbuds, in particular, presents social acceptance challenges for DHH individuals, making such systems impractical for everyday use.

To address all the above critical limitations, it is imperative to develop effective, user-friendly, and practical solutions that can comprehensively capture all ASL components and translate ASL

into English accurately. One promising direction lies in leveraging advances in wearable camera technologies. Recently, with the rapid development of deep learning in computer vision, embedded wearable cameras have become increasingly popular in studies of human behavior in real-world settings, offering a powerful means of visually capturing actions. In particular, egocentric (first-person view) cameras, which align with the wearer's field of view, have shown great potential for capturing nuanced human actions and interactions in natural environments [7, 16, 17, 27]. Moreover, the rapid advancement and widespread availability of smart glasses, such as Google AR glasses [28, 36] and Meta AI glasses [27], underscore their growing usability, making them increasingly practical for seamless integration into everyday life.

Motivated by these technological advancements and practical opportunities, we propose an ASL recognition and translation system that utilizes three wearable cameras mounted on a single pair of smart glasses. Additionally, our user experience survey study shows that SignGlass was favored by users over existing alternative devices, including earphones, smart rings, and smartwatches. More usability study is discussed in Sec. 3.4. This system is designed to simultaneously capture both manual and non-manual markers, enabling robust, accurate, and comprehensive ASL translation across diverse users and environments. In summary, as shown in the left side of Fig.1, we utilize a pair of glasses with three wearable cameras to capture the manual and non-manual markers. The middle camera will record hand gestures to recognize different ASL manual markers. The two side cameras will be used to identify different non-manual markers, including facial motions and the view changes generated by head motions. After recording these two essential components, as shown on the right side of Fig.1, we designed advanced deep learning algorithms to achieve high performance of cross-user ASL recognition and translation. The contributions of our work can be summarized as follows:

- To the best of our knowledge, SignGlass is the first single smart wearable device for comprehensive ASL recognition and translation. We developed and implemented a portable prototype smart glass equipped with three wearable cameras to recognize both manual and non-manual markers simultaneously and convert ASL glosses into English. A user survey with 46 Deaf participants revealed that 39 expressed willingness to use our SignGlass system for daily communication.
- We propose an end-to-end deep learning algorithm with multi-stage optimization to achieve high ASL recognition and translation performance. We first propose a jitter-aware spatio-temporal encoder designed for stabilizing first-person view hand gestures. We then introduce a signer-conditioned FiLM module that leverages facial dynamics to enable user-independent modeling of non-manual markers. Lastly, we present a lightweight gloss-to-English translation layer that supports grammatical ASL-to-English mapping using multi-modal gloss inputs.
- We adopted a novel cascade data augmentation method to accommodate the ASL sign diversity across different ASL users' preferences and styles.
- We conduct an extensive user study with native sign language users ( $N = 14$ ) to validate the feasibility of SignGlass.

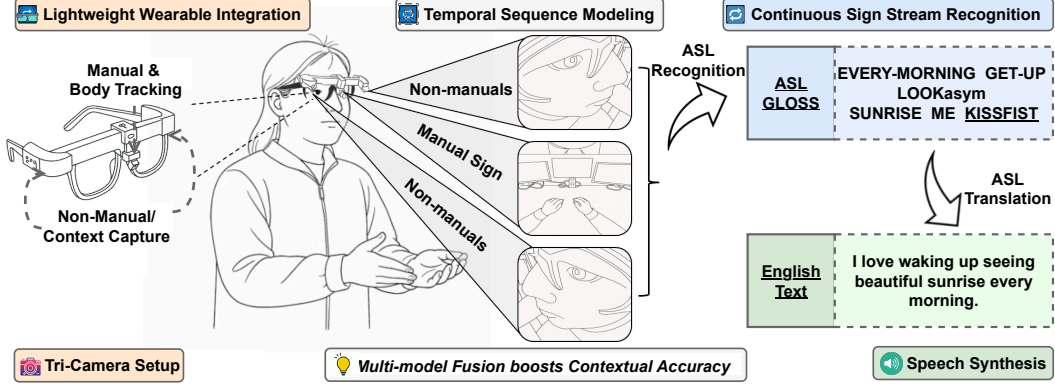


**Figure 1: Overall flow of SignGlass. We designed the SignGlass hardware system and advanced effective deep learning algorithms to implement accurate cross-user ASL recognition and translation.**

Our system achieved a WER of 16.98% and a BLEU-1 score of 83% in leave-one-user-out testing, showing strong generalizability.

## 2 RELATED WORK

Table 1 contrasts SignGlass in the context of key-related works. Among this table, the accuracy metric *WER under UI* represents word error rate (WER) under user-independent (UI), which represents the ASL capture accuracy, defined in Section 6.3. The *BLEU score under UI* represents the calculation of the Bilingual Evaluation Understudy (BLEU) score under user-independent (UI), which evaluates machine-translated text quality, defined in Section 6.3. *Non-manual Marker* refers to recognizing non-manual markers (e.g., facial expressions and head motion). *Unseen User Support* refers to recognizing ASL patterns in new users. *Deaf User Study* indicates if the study was conducted with deaf users. Then, this section elaborates on manual markers recognition, non-manual markers recognition, and traditional camera-based ASL recognition.

### 2.1 Manual Markers Recognition in ASL

Wearable sensor-based ASL recognition has attracted significant attention due to its portability, making it ideal for daily use. Various sensors, including IMU, acoustic sensors, and EMG (Electromyography), have been extensively explored.

IMU sensors, worn on the hands, fingers, or wrists [45, 49, 76, 80], capture hand motions, while acoustic sensors can be integrated into devices like earphones [40]. EMG sensors [76, 77] detect muscle activity associated with finger and hand movements.

Despite their practicality, these technologies often have limited effectiveness in user-independent scenarios or when recognizing predetermined gestures. For example, Ekiz et al. [25] recognized only 13 sentences with a smartwatch, whereas SignSpeaker [31] recognized 73 sentences under similar conditions. Systems such as SonicASL [40] and TransASL [37] demonstrated effectiveness only in user-dependent environments.

In contrast, SignGlass utilizes wearable cameras to comprehensively capture and interpret gestures from previously unseen users. Additionally, our user study with Deaf participants (see Section 3.4) revealed that hand-based wearable devices were perceived as restrictive and intrusive. Unlike prior work, SignGlass introduces

an innovative approach from the perspective of DHH individuals by employing smart glasses with embedded cameras. This hands-free design enhances comfort for prolonged daily use while improving ASL recognition accuracy and user-friendliness.

### 2.2 Non-manual Markers Recognition in ASL

Non-manual markers in ASL, such as facial expressions, head movements, and body postures, are crucial because they convey grammatical information, including tone, questions, and negation, thus enriching the signed message [63, 69]. These visual cues complement manual signs to clarify meaning, express emotions, and differentiate sentence structures. Recognizing non-manual markers is more challenging compared to manual markers due to their often subtle, context-dependent, and fine-grained movements.

We observed a growing trend toward recognizing non-manual markers in sign language, including eyebrow and head gesture tracking [47], adaptive 2D-3D face tracking [53], and tracking via wearable devices [54]. Recent works, such as TransASL [37] and SmartASL [41], proposed using in-ear sensing to detect facial expressions. However, these studies primarily capture ear canal deformations, limiting performance and practical use in the DHH community. Conversely, SignGlass incorporates wearable cameras embedded within smart glasses, significantly enhancing performance while providing adequate portability and mobility.

### 2.3 Sign Language Recognition using Cameras

Sign language recognition has primarily leveraged vision-based techniques, aligning manual and non-manual markers with extracted visual features [8, 29, 30, 55, 81]. Approaches such as SignBERT [33] and HMA [32] decode 3D hand keypoints from RGB videos. Camgoz et al. [20] introduced a transformer-based architecture that simultaneously performs continuous sign language recognition and translation using Connectionist Temporal Classification (CTC) loss.

Despite demonstrating promising results for recognizing both manual and non-manual markers, these systems typically lack portability, restricting practical daily use. Recent efforts include Zhao’s smart-glasses-based system [78], designed to translate sign language captured from a camera facing the user, and Sign-to-911

**Table 1: Summary of Main Related Work.** UI represents User-independence. WER represents Word Error Rate. BLEU represents Bilingual Evaluation Understudy.

| Metrics                    | SignGlass | SignRing [45] | WearSign [76]   | ASLRing [80] | SmartASL [41]   | TransASL [37]   |
|----------------------------|-----------|---------------|-----------------|--------------|-----------------|-----------------|
| <b>Non-manual Marker</b>   | Yes       | No            | No              | No           | Yes             | Yes             |
| <b>No. of Devices</b>      | 1         | 2             | 2               | Multiple     | 2               | 2               |
| <b>Form Factor</b>         | Glass     | Rings         | Watch + Armband | Rings        | Earbuds + Watch | Earbuds + Glass |
| <b>Sensor</b>              | Camera    | IMU           | IMU + EMG       | IMU          | IMU + Mic       | Mic             |
| <b>ASL Translation</b>     | Yes       | No            | Yes             | No           | Yes             | No              |
| <b>Unseen User Support</b> | Yes       | No            | Yes             | Yes          | No              | No              |
| <b>No. of Sentences</b>    | 600       | 951           | 250             | 1080         | 40              | 40              |
| <b>No. of Vocabulary</b>   | 1052      | 934           | 56              | 1057         | 80              | 80              |
| <b>WER under UI (%)</b>    | 16.98     | No Tracking   | 4.7             | 26.9         | ≈75             | ≈84             |
| <b>BLEU under UI (%)</b>   | 83        | No Tracking   | No Tracking     | No Tracking  | No Tracking     | No Tracking     |
| <b>Deaf User Study</b>     | Yes       | No            | No              | Yes          | No              | No              |

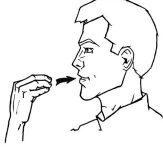

**Figure 2: ASL sign: Eat [5]**

**Figure 3: ASL sign: Bike [4]**

[29], employing AR glasses for bidirectional ASL-English translation in emergency scenarios. However, these systems primarily rely on front-facing cameras or focus exclusively on non-manual markers. In contrast, SignGlass employs portable smart glasses embedded with lightweight cameras to recognize both manual and non-manual markers without relying on front-facing cameras, enhancing practicality and portability for everyday use.

### 3 DESIGN CONSIDERATIONS

In this section, we begin with an introduction to the basics of American Sign Language (ASL), followed by a discussion of system design requirements, performance considerations, and user experience study.

#### 3.1 Basics of American Sign Language (ASL)

Sign language is a visual language characterized by intricate gestures and grammatical structures. It comprises both manual markers (e.g., hand shape, movement, palm orientation, and location) and non-manual markers (including facial expressions, head tilts, and body positions) [15, 42, 67, 72, 73, 77]. Sign language is considered a class of natural language, possessing its own distinctive grammar and lexicon. There are over 200 sign languages with millions of people, indicating the necessity of translation [6]. The majority of sign language used in the USA and parts of Canada is ASL.

**(i) Manual and Non-Manual Markers.** As the central part of ASL, Manual markers include four basic parameters—hand shapes, positions, and movements—play [37]. Each manual marker is uniquely crafted to convey specific meanings by positioning and moving the hands with distinct hand shapes and palm orientations. Figures 2 and 3 depict the hand movements necessary for signing the words

"eat" and "bike". Accurate identification and differentiation of manual markers are essential for precise ASL word recognition and interpretation, as they each contribute critical meaning to a signed message.


**Figure 4: Examples of ASL grammatical elements from multimodalities including facial expressions, head movements, and hand gestures: WH-questions (WH-Q), Yes-No questions (Y/N), Negation (NEG), Topicalization (TOP), Conditional sentences (C/W) and Others sentences (Others) from ASLLRP[9].**

As shown in Figure 4, Non-manual markers (NMMs), including non-affective facial expressions, head tilting, and body positioning, are essential in ASL for conveying grammatical nuances and providing semantic context to manual markers (e.g., hand shapes and movements) [52]. NMMs enhance comprehension by differentiating sentence types and adding necessary context. For example, it is necessary to construct the following five primary sentence types utilizing NMMs: 1). *WH-questions*: Furrowed eyebrows and head tilts signal questions that request specific information. 2). *Yes-No questions*: Raised eyebrows and forward head movements indicate a yes-no inquiry. 3). *Negation*: Head shakes and negative facial expressions denote denial. 4). *Topicalization*: Raised eyebrows and pauses emphasize or clarify the topic. 5). *Conditional sentences*: Raised eyebrows and pauses articulate conditional clauses. Unfortunately, current wearable ASL recognition technologies will add another wearable device - earbuds - to sense facial expressions by detecting ear canal deformations[37, 41] using IMU sensing or acoustic

sensing. However, these earbuds-based non-manual markers or facial expression detection are highly restricted by generalizability because the leave-one-user-out performance is low. To overcome these limitations, we propose a novel system capable of recognizing both manual and non-manual markers simultaneously, enabling a more comprehensive and accurate interpretation of ASL.

**Table 2: ASL sentence vs English sentence**

| ASL Sentence | English Sentence           |
|--------------|----------------------------|
| LOVE YOU     | I love you.                |
| WRONG WHAT   | What’s wrong?              |
| REMEMBER ME  | Do you remember me?        |
| ME SICK      | I’m not feeling very well. |

(ii) **ASL Grammar** After recognizing the manual and non-manual markers, we will get the ASL glosses conforming with unique linguistic rules and structures. ASL sentences primarily follow the *Subject-Verb-Object* structure [10]. However, the order can shift when the topic is introduced first—a process known as topicalization [61]—resulting in the format: *Object, Subject-Verb*. Table 2 presents several examples comparing ASL sentence structures with their English translations.

## 3.2 System Design Requirements

(i) **ASL Differences Across Different Users.** Similar to English, ASL signers will exhibit a diverse and nuanced range of variations based on their preferences and styles, influenced by regional, cultural, and community influences[80]. This diversity, especially the more complex co-articulated signs used by native signers compared to learners, presents significant challenges for recognition tasks in ASL interpretation. Prior works [37, 41] focus on the user-dependent study, which means those systems can only recognize their own collected data without any contribution. Moreover, the joints of various fingers exhibit intricate inter-dependencies that are not readily captured by equations; however, we anticipate that our ML models will inherently discern these constraints.

(ii) **Unseen Sentence Generalization.** In real-world applications, ASL recognition systems must cope not only with inter-signer variability but also with a vast space of previously unseen sentence structures. Unlike fixed-vocabulary gesture recognition systems, ASL exhibits a highly compositional grammar where a finite set of glosses can be combined in diverse ways to express new meanings [12]. For example, the glosses *LEARN*, *SIGN*, and *WANT* may appear in multiple valid permutations depending on emphasis or discourse style (e.g., “LEARN SIGN WANT” vs. “WANT LEARN SIGN”). These permutations lead to a long-tail distribution of possible sentence constructions that pose significant challenges for generalization. Traditional ASL recognition pipelines often rely on scripted phrase datasets or predefined template matching, which limits their ability to handle novel combinations [13, 43]. Moreover, many prior studies evaluate models under user-dependent settings, where sentence compositions in training and testing closely resemble each other. Such setups fail to reflect the variability encountered in open-domain ASL usage.

To address this, we explicitly design our system for unseen sentence generalization by leveraging user-independent training, multimodal gloss construction, and translation-aware decoding. Our translation module conditions output generation on the full gloss context without assuming fixed phrase structures [65]. As demonstrated in Section 6, the system maintains robust performance even under leave-one-user-out testing with unseen sentence arrangements.

## 3.3 Performance Considerations

(i) **Effective ASL Sensing Using a Single Device.** Considering that non-manual markers are primarily situated around the head and face, and manual markers are located near the chest area, using wearable devices to monitor them at the same time is challenging. Currently, TransASL[37] and SmartASL[41] have started to explore the possibility of sensing manual and non-manual markers using a combination of glass and earbuds, or a combination of smartwatch and earbuds. This combination will restrict practical usage in daily life for two reasons, one is that the earbud is not useful for deaf people at all, and the other is that specific multi-device will limit the deployment in daily life. Our goal is to provide a single comfortable-wearing device while concurrently preserving the capacity to detect both manual and non-manual markers.

(ii) **Field-of-View Considerations.** Our proposed system uses a smart glasses camera with an ultra-wide 160° field-of-view (FOV) to ensure comprehensive coverage of both manual and non-manual markers during ASL signing. This wide FOV allows the entire upper body, including hand gestures and facial expressions, to remain within the frame, reducing the risk of missing key ASL features. Although such wide-angle lenses introduce notable image distortion (e.g., fisheye effect), this is mitigated through standard calibration techniques (e.g., OpenCV fisheye correction) and does not require real-time correction during operation. In practice, this trade-off allows broader spatial coverage and helps to keep data naturally centered across various gestures without additional alignment processes.

(iii) **Mutli-channel and Multi-modal Recognition based on Deep Learning.** After capturing manual and non-manual markers, recognizing them accurately is unavoidable. To accurately recognize ASL manual markers and non-manual markers, a multi-channel and multimodal recognition system is demanded. In addition, the restructuring and omission in translating English to ASL glosses highlight a critical point: ASL uses a different syntactic structure and prioritizes different linguistic elements. This example shows how ASL glosses omit certain English grammatical elements while retaining the core meaning, showcasing ASL’s unique linguistic framework. This difference poses a significant challenge for hearing individuals trying to understand ASL without a deep understanding of its grammar and nuances. It underscores the importance of comprehensive education and immersion in ASL for effective communication within the Deaf community, illustrating the richness and depth of ASL as its language. After processing the sensing data, we complete the ASL recognition, preserving the uniqueness of ASL glosses.

(iv) **Generalizability Across Various Users.** Due to the vast differences among individuals, even for the same gesture, a significant decline in the accuracy of gesture recognition for new users [40, 45, 68]. In our work, we proposed an optimized algorithm for

data augmentation, which is capable of duplicating and reinforcing key gestures within the video sequence to make them more prominent, potentially improving the accuracy and robustness of gesture recognition.

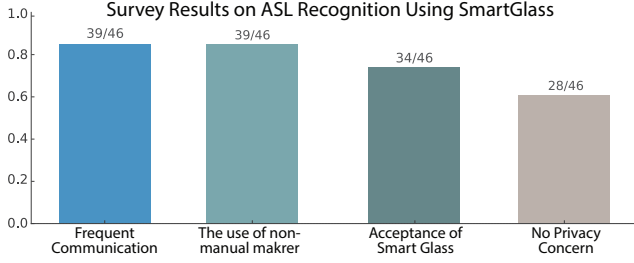


Figure 5: Survey Results on ASL and SignGlass Usage

### 3.4 User Experience Study

**(i) User-Centered Design Insights from DHH Community Interviews.** To design a human-centered system for sign language translation tailored specifically to the needs of the Deaf community, we conducted an in-person interview study involving 46 participants from the Deaf and Hard-of-Hearing (DHH) community. This study focused on four key metrics to inform the system’s design. First, frequent communication assesses how often deaf individuals interact with hearing people, which is essential for determining the real-world utility of a translation system. Second, the use of Non-Manual Markers examined how deaf individuals incorporate facial expressions and other non-manual cues in their signing, highlighting the need for comprehensive recognition beyond hand gestures. Third, the Acceptance of Smart Glasses explored how comfortable participants felt wearing smart glasses in public. Given concerns that such devices might draw unwanted attention and lead to feelings of social exclusion during signing, we addressed this issue within the context of our interview-based user study. Fourth, Privacy Concerns of Smart Glasses investigated participants’ apprehensions about potential invasions of privacy when using wearable technology. As shown in Fig. 5, 39 participants reported regularly communicating with hearing individuals and using non-manual markers, reinforcing the importance of inclusive translation technology that supports both groups. Additionally, the majority of participants were receptive to using smart glasses, with 28 expressing little or no concern about privacy. Drawing from these insights, we propose the development of smart glasses integrated with cameras capable of recognizing and translating both manual and non-manual ASL markers, thereby supporting seamless and respectful communication across communities.

**(ii) User Experience Survey Across Various Wearable Platforms.** We conducted a usability survey to evaluate the user experience of the SignGlass sensor device in comparison to three ASL recognition devices: SmartASL [41], TransASL [37], and ASLRing [80]. Afterward, they rated each device in terms of *ease of use*, *comfort*, and *appearance* on a scale from 0 to 10. The results, shown in Fig. 6, indicate that TransASL was less favorable due to the wearable devices with complex wires limiting the usage of TransASL. The SmartASL and ASLRing provided enhanced comfort but necessitated the use of several devices to monitor both manual and

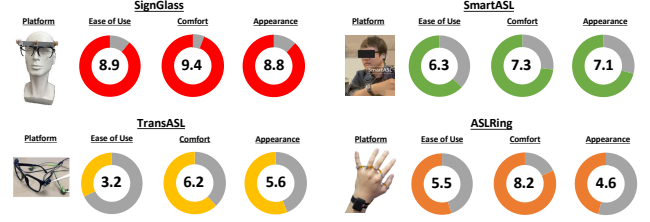


Figure 6: User experience survey on SignGlass compared with three alternative sensing devices.

non-manual markers, leading to some inconvenience. Fig. 6 shows that our current SignGlass work stands out for its usability during daily activities due to the highest scores on usability, comfort, and appearance.

## 4 HARDWARE DESIGN

This section outlines the design and implementation details of our proposed SignGlass platform.

### 4.1 Choice of Sensing Modality

Recent advances in ASL recognition have explored a range of sensing modalities mounted on wearable devices, including acoustic sensing [37, 40, 46], electromyography (EMG) [76, 77], and inertial measurement units (IMUs) [41, 45, 79, 80]. However, these approaches face notable limitations in fully capturing the manual and non-manual richness of ASL. (i) Acoustic sensing via earbuds leverages Doppler shifts or ear canal deformations for single-word ASL recognition [41]. While feasible, these methods are limited by small vocabularies and cannot detect crucial non-manual markers like facial expressions or head movements. (ii) EMG sensors on the wrist or forearm capture muscle signals for fine-grained hand gestures [50, 76, 77] but require user-specific calibration and precise placement. They also fail to capture non-manual cues without adding intrusive facial sensors. (iii) IMUs, used in wearables like smartwatches and rings [45, 49, 80], support dynamic gestures but struggle with static hand shapes and facial grammar. They require calibration, are prone to drift, and often need multiple devices, limiting daily usability.

In contrast, cameras offer comprehensive and high-resolution visual data that can capture detailed hand gestures, body gestures, and fine-grained facial expressions. They eliminate the need for calibration, scale well with large datasets [24], and support robust machine learning models without extensive per-user adaptation. Thus, we propose a multi-camera system, SignGlass, for sign language recognition and translation that simultaneously tracks fine-grained manual markers (e.g., 3D finger motion and hand pose) and non-manual markers (e.g., facial expressions and head movement).

### 4.2 Justification on Form Factor Design.

As illustrated in Fig. 7, the objective of our design is to create a discreet, adaptable, and comfortable smart glass that can be worn continuously throughout the day and night. The device is capable of sensing hand gestures, facial expressions, and other non-manual

markers while wirelessly streaming data for real-time analysis. Existing platforms, such as sensor gloves [1] and strap-based devices [3], restrict natural and dexterous finger movements and are unable to detect non-manual markers like facial expressions, which are critical for American Sign Language (ASL) recognition. Although systems like SmartASL [41] and TransASL [37] support both manual and non-manual markers, their reliance on earbuds—a device often redundant for Deaf and Hard of Hearing (DHH) individuals—limits their comfort and usability. Previous studies [62] have highlighted the challenges of earbud-based technologies for DHH users, as they may interfere with other assistive devices, such as hearing aids or cochlear implants, further reducing practicality. Our proposed solution—a smart glass with three mounted cameras—offers a more accessible and inclusive form factor by eliminating the need for earbuds. In addition, recent developments such as Google AR glasses [36], which transcribe spoken language and display it on AR lenses in real time [28], demonstrate the potential of this technology to enhance communication for DHH individuals by enabling them to read transcribed speech. This innovation supports the broader goal of fostering two-way communication between DHH individuals and hearing people, further bridging the communication gap.

To optimize both functionality and user comfort, we focused on the following design features: (1) SignGlass integrates three miniature RGB cameras powered by a Raspberry Pi Zero to ensure compactness and mobility. (2) The cameras are enclosed in a custom 3D-printed case made from Thermoplastic Polyurethane (TPU) using fused deposition modeling (FDM), ensuring comfort for skin contact and suitability for long-term wear.

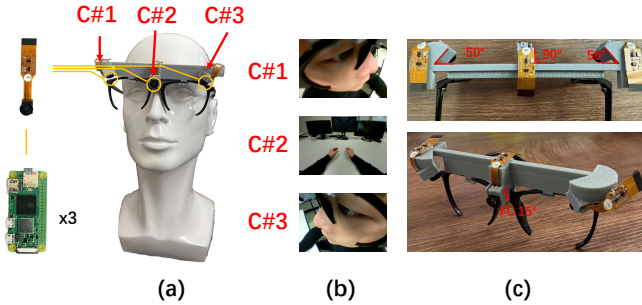


**Figure 7: SignGlass experimental prototype on 3D head model.** While users are performing the sign language, the cameras are capturing images of facial movement and hand motion. (a) SignGlass platform: data collection module: three miniature RGB cameras, three Raspberry Pi Zero and a 3D-printed case. (b) Images from the three cameras, including C#1, C#2, C#3. (c) Camera angle settings. The angle settings are labeled with FC values, which refer to the front camera (C#2) angle. The detailed camera settings are included in Section 4.4.

### 4.3 Hardware

To meet the hardware design considerations outlined in Section 3.3, the SignGlass prototype incorporates three miniature RGB cameras (i.e., OV5647 by Arducam [2]), each with a fixed field of view (FOV) of 160° and dimensions of 11.5 mm × 9.5 mm. As depicted in Fig.

7, the cameras are strategically positioned to capture both facial expressions and hand movements from multiple angles. Further details regarding the arrangement of the camera positions are provided in Section 4.4. The cameras capture images at a resolution of 720 × 480 pixels at 30 frames per second (FPS), and the collected images are transmitted wirelessly to a mobile device via WiFi. Each camera is connected to a Raspberry Pi Zero through a cable, and the entire platform is powered by an external power bank. The Raspberry Pi Zero and the battery are housed within a custom-made case fabricated from PLA material. The total weight of the SignGlass platform is 40 grams.

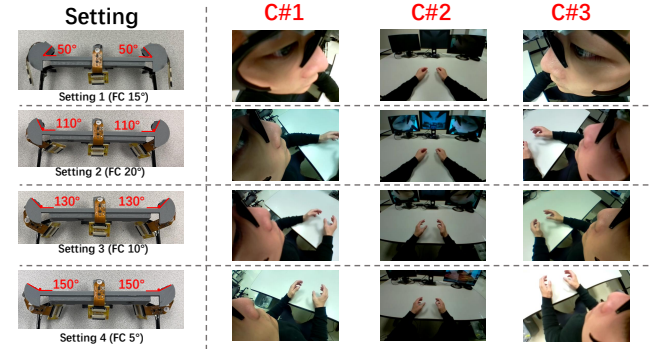


**Figure 8: Impact of camera angles on field of view.** This figure demonstrates how different camera angles influence the information captured by each camera (C#1, C#2, and C#3). As the angles of cameras C#1 and C#3 increase, the amount of facial information in their respective views decreases. Camera C#2, with its consistent front-facing overhead angle, provides a stable top-down perspective across settings.

### 4.4 Camera Setting

Fig. 7 presents the images captured by three distinct cameras integrated into SignGlass. To determine the optimal camera settings, we experimented with the placement of the cameras in four different settings as demonstrated in Fig. 8. An optimized camera configuration (in terms of number, position, and orientation) is crucial for accurately estimating facial expressions and hand movements. The analysis of the impact of cameras is discussed in Section 6.3.

We evaluated the camera positioning by considering (1) non-manual markers, including facial expressions, head movements, and body posture, (2) the range of hand movements during sign language, and (3) the camera's field of view (FOV) of 160°. Consequently, we empirically positioned the camera (see Fig. 7 C#2) in the middle of the glasses at 15° perpendicular to the frame. This ensures that SignGlass can effectively capture hand motion data. The other two cameras (see Fig. 7 C#1 and C#3) are positioned on the sides of the glasses' legs to focus on non-manual markers, such as facial expressions. The positions and arrangements were empirically determined based on preliminary experiments conducted by the researchers. The rotation angles and tilt of the cameras are depicted in Fig. 8 (tilt in blue and rotation in red). The purpose of tilting and rotating cameras C#1 and C#3 is to better orient them toward the face.

## 5 ALGORITHM DESIGN

In this section, we present the algorithmic architecture of SignGlass. As illustrated in Fig. 9, SignGlass comprises two main components: ASL-to-Gloss recognition and Gloss-to-English translation. EgoCorrNet processes first-person RGB and motion data to extract manual gloss tokens, while EgoFaceNet performs dual-stream temporal encoding and FiLM-conditioned fusion to classify non-manual glosses. The outputs from both modules are concatenated into a structured gloss sequence, which is then decoded into natural English using a Transformer decoder trained on gloss-English pairs. This modular pipeline enables interpretable, real-time ASL translation with explicit linguistic grounding.

### 5.1 ASL-to-Gloss Implementation

#### 5.1.1 Manual Marker Recognition.

First-person (egocentric) video capture introduces unique noise patterns due to natural head motion, self-occlusion, and background distractions, which significantly degrade gesture recognition accuracy. In contrast, prior third-person approaches are often overfitted to fixed camera setups and fail to generalize in wearable, mobile scenarios. To overcome this, we introduce **EgoCorrNet**, a jitter-aware, spatio-temporal recognition framework designed specifically for first-person ASL input. EgoCorrNet translates first-person RGB video into symbolic ASL gloss sequences, serving as the manual recognition stream in our overall system (Fig. 9).

The input to EgoCorrNet consists of two complementary motion cues: dense optical flow and hand landmarks. While MediaPipe [75] effectively extracts 21-point hand landmarks per frame, this representation alone is unreliable in the presence of occlusions or camera-induced jitter. We find that optical flow trajectories provide critical temporal context, especially for capturing continuous hand movement during complex gestures. These dual-channel inputs are spatially aligned and fused early at the channel level, allowing for lightweight downstream processing and reduced inference latency. Formally, the input at time  $t$  is represented as a fused tensor  $\mathbf{x}_t = [\mathbf{f}_t; \mathbf{l}_t] \in \mathbb{R}^{H \times W \times C}$ , where  $\mathbf{f}_t$  is the RGB optical flow map and  $\mathbf{l}_t$  is a 21-joint hand landmark heatmap rendered as a spatial feature map. This fused multi-modal representation forms the input to our spatio-temporal encoder, which we describe next. Additionally, a binary motion mask is computed from the optical flow magnitude and used to suppress background clutter during feature extraction.

The spatio-temporal encoder is processed using the CorrNet+ [34] architecture as a backbone, which combines a MobileNet encoder with squeeze-and-excitation (SE) attention for compact spatial modeling, followed by a Bi-LSTM for capturing local temporal dependencies. We build on this foundation by introducing a jitter-aware cross-modal alignment module to address the instability unique to egocentric video, such as head-induced jitter and self-occlusion. To robustly align motion cues under egocentric jitter, we propose a jitter-aware cross-modal alignment module that dynamically links optical flow and landmark sequences over time. Given temporal feature sequences from both modalities, we compute bidirectional attention to align their latent representations:

$$\mathbf{Z}_F = \text{softmax} \left( \frac{(\mathbf{F}W_Q)(\mathbf{L}W_K)^T}{\sqrt{d}} \right) (\mathbf{L}W_V)$$

$$\mathbf{Z}_L = \text{softmax} \left( \frac{(\mathbf{L}W_Q)(\mathbf{F}W_K)^T}{\sqrt{d}} \right) (\mathbf{F}W_V)$$

Here,  $\mathbf{F}, \mathbf{L} \in \mathbb{R}^{T \times d}$  denote flow and landmark features over  $T$  frames, and  $W_Q, W_K, W_V$  are learned projection matrices. These attention maps allow each modality to re-weight and adapt to the other, compensating for temporal inconsistencies such as motion blur in flow or missing joints in landmarks. We further stabilize alignment using residual gating to fuse  $\mathbf{Z}_F$  and  $\mathbf{Z}_L$  into a unified, jitter-insensitive alignment for temporal consistency representation. The aligned multimodal sequence is decoded into ASL gloss tokens using a lightweight temporal decoder comprising a 1D convolutional layer followed by a BiLSTM. The 1D CNN captures local motion patterns, while the BiLSTM models long-range dependencies across the sequence. A fully connected classification layer is attached on top to predict gloss distributions for each frame.

We adopt a Connectionist Temporal Classification (CTC) loss to supervise this process, enabling alignment between unsegmented input video frames and gloss sequences without requiring frame-level annotations. This gloss-aware decoding strategy accommodates variable-length signing streams while maintaining low-latency inference. It forms the basis for downstream interactive features such as correction feedback and sign replay within our tutoring system.

#### 5.1.2 Non-Manual Markers Recognition.

Non-manual markers (NMMs)—including facial expressions, mouth morphemes, eye gaze, and head posture—play a central role in encoding sentence-level grammar. These cues often indicate question types, negation, or topical structures and may overlap in time. However, existing systems frequently omit non-manual modeling due to occlusion, asymmetric viewpoint, and annotation difficulty.

To address this gap, we introduce *EgoFaceNet*, a dual-camera, temporally-aware module for classifying sentence-level grammatical categories from stereo facial video. EgoFaceNet is designed to detect multiple overlapping non-manual cues and incorporate contextual conditioning from manual gloss sequences. It outputs a 4-dimensional binary vector  $\mathbf{p} \in [0, 1]^4$ , indicating the presence of four grammatical types: WH-questions (WH-Q), Yes/No questions (Y/N), Topic/Comment structures (T/C), and Negation (NEG). EgoFaceNet processes dual-view facial video captured by synchronized left and right cameras, denoted as  $\mathbf{I}^L, \mathbf{I}^R \in \mathbb{R}^{T_b \times H \times W \times 3}$ . Each RGB stream is temporally sampled at 30 fps and spatially resized to  $224 \times 224$ , followed by per-pixel normalization. The video frames are partitioned into non-overlapping spatial patches of size  $16 \times 16$ , resulting in  $T = 64$  patches per view. Each patch is linearly projected into an embedding vector and enriched with positional encoding to form token sequences for temporal modeling.

To capture fine-grained non-manual motion, we require a spatio-temporal encoder that models temporally localized expressions—such as eye blinks, gaze shifts or brief eyebrow raises—without sacrificing spatial detail. We adopt TimeSFormer [11], which factorizes spatial and temporal attention into separate blocks within each transformer layer. This design enables high-resolution attention across both dimensions while remaining computationally efficient.

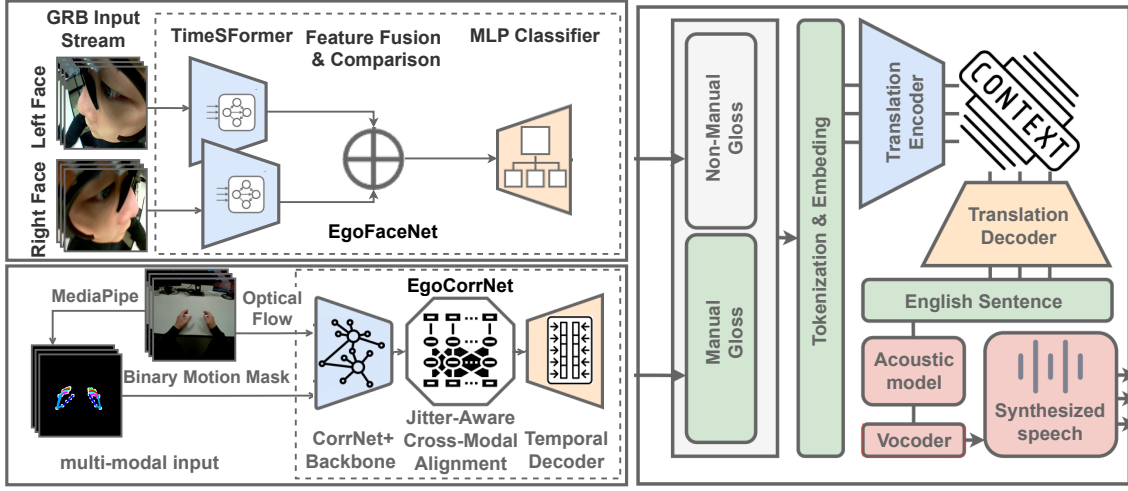


Figure 9: A flow diagram shows the processing mechanism of SignGlass, which consists of the ASL videos-to-ASL Gloss recognition module and the ASL Gloss-to-English translation module.

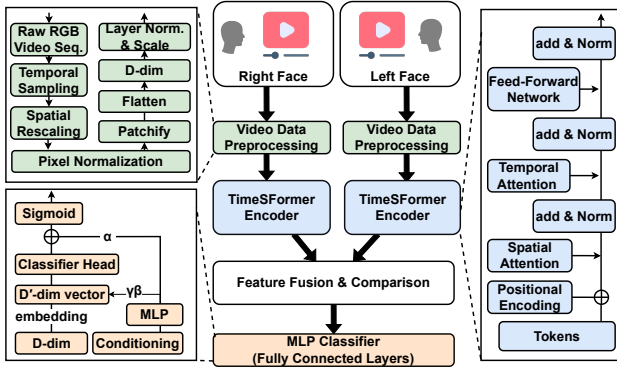


Figure 10: Overview of the proposed EgoFaceNet for facial type classification. The architecture includes two parallel TimeSFormer encoders for left and right facial video inputs, followed by feature fusion, FiLM-based conditioning, and a multi-label classifier head.

and is especially effective for modeling short-term facial dynamics embedded in continuous sign sequences.

We apply two independent TimeSFormer encoders to the left and right view token sequences, yielding outputs  $F^L, F^R \in \mathbb{R}^{T \times D}$  with  $D = 768$ . Each encoder contains 6 transformer blocks with 8 attention heads and feed-forward sublayers. The encoders do not share weights, allowing them to specialize in asymmetric view-dependent features such as partial occlusion or illumination imbalance. Temporal alignment is preserved through synchronized capture, both streams retain frame-wise correspondence for subsequent fusion. While each view captures complementary facial signals, reliable grammatical inference requires integrating both perspectives and grounding them in the linguistic context of the signed utterance. To achieve this, we fuse the encoded features from both views and apply contextual conditioning using gloss-level embeddings from the manual stream. Specifically, we concatenate the TimeSFormer

outputs  $F^L, F^R \in \mathbb{R}^{T \times D}$  along the channel dimension to obtain fused tokens  $F^{\text{cat}} \in \mathbb{R}^{T \times 2D}$ . These are passed through a 2-layer residual MLP with GELU activation, a hidden dimension of  $2D$ , and output projection back to  $D$ , applied frame-wise. The residual connection is added after the second layer, and layer normalization is applied before the first. This produces temporally aligned fused features  $F^{\text{fused}} \in \mathbb{R}^{T \times D}$ .

To incorporate semantic grounding, we condition these features using FiLM [60], modulated by gloss context. We extract token embeddings  $\{g_i\}_{i=1}^N \in \mathbb{R}^{N \times D'}$  from the manual-stream decoder (see Section 5.1.1) and mean-pool them into a sentence-level vector  $g_{\text{avg}} \in \mathbb{R}^{D'}$ . Two linear projections map  $g_{\text{avg}}$  to scale and shift vectors  $\gamma, \beta \in \mathbb{R}^D$ , which are broadcast across time and used to modulate the fused sequence:

$$H_{\text{FiLM}}(t) = \gamma \cdot F^{\text{fused}}(t) + \beta, \quad \forall t \in [1, T]$$

This results in context-aware feature representations  $H_{\text{FiLM}} \in \mathbb{R}^{T \times D}$  that reflect both visual dynamics and concurrent manual semantics. Notably, FiLM is applied at the feature level without requiring explicit temporal alignment between gloss tokens and facial frames. The context-conditioned feature sequence  $H_{\text{FiLM}} \in \mathbb{R}^{T \times D}$  is mean-pooled along the temporal axis to obtain a sentence-level embedding  $h_{\text{sent}} \in \mathbb{R}^D$ . This vector is passed through a 3-layer MLP classifier with hidden dimensions [256, 128, 4], ReLU activations, and a final sigmoid layer to produce a 4-dimensional probability vector  $\mathbf{p} \in [0, 1]^4$ . Each dimension  $p_k$  corresponds to the likelihood of one grammatical category being present: WH-Q, Y/N, T/C, or NEG. The model is trained using multi-label binary cross-entropy loss:

$$\mathcal{L}_{\text{grammar}} = - \sum_{k=1}^4 [y_k \log p_k + (1 - y_k) \log (1 - p_k)]$$

where  $y_k \in \{0, 1\}$  is the ground-truth label for class  $k$ . Class weights are not applied, as the dataset is approximately balanced across categories.

We adopt sentence-level supervision because annotating precise frame- or segment-level labels for non-manual markers is inherently ambiguous and subject to low inter-rater agreement, especially for transient or blended expressions. By using binary sentence-level tags, we provide a stable and reproducible training signal while still capturing the presence of overlapping grammatical features.

The dual-view design improves robustness to egocentric variability such as facial occlusion by the hands, lateral head tilt, asymmetric lighting, or camera motion. Even when one side of the face is partially occluded or overexposed, the opposing view often retains key expression features like mouth shapes or eyebrow positions. This redundancy is essential in first-person scenarios where extreme viewpoint shifts and environmental clutter are common.

In ablation experiments (Section 6.6.8), EgoFaceNet improves the grammatical classification F1 score by 5.8% over a single-view TimeSFormer baseline without FiLM conditioning. The gain is statistically significant ( $p < 0.01$ , paired t-test across five folds), validating the complementary contributions of multi-view fusion and semantic conditioning. As shown in Fig. 9, EgoFaceNet constitutes the non-manual recognition stream in our system, enabling sentence-level grammatical analysis and supporting grammar-aware feedback generation.

## 5.2 Gloss-to-English Translation

To train the gloss-to-English translation module, we constructed a paired dataset consisting of ASL gloss sequences aligned with natural English sentences. The training data was compiled from two sources: (1) ASLLRP Sign Bank [9], a linguistically annotated ASL corpus with gloss-level transcripts, and (2) a custom-aligned subset of the How2Sign dataset [24], where we manually aligned gloss sequences to reference English translations with the assistance of professional interpreters. In total, the final training set consists of 11,650 sentence pairs covering diverse ASL grammatical categories such as topicalization, WH-questions, and conditional statements. We applied Byte-Pair Encoding (BPE) for subword tokenization on both gloss and English tokens to handle vocabulary sparsity. Alignment was conducted under a monotonic but not strictly one-to-one assumption, allowing the model to learn common reordering patterns in ASL grammar (e.g., WH-element fronting, topicalized object placement).

To accommodate the structural differences between ASL and English, we employed a Transformer-based decoder trained with label smoothing ( $\epsilon = 0.1$ ) and beam search (beam size = 4). This setup enables fluent sentence generation while supporting grammatical reordering and disambiguation.

Beyond BLEU scores, we conducted a qualitative analysis of predictions. For example, the gloss *“FAVORITE MOVIE CAPTION TURN ON”* is fluently translated as *“I like to watch movies with the subtitles on.”*, successfully capturing subject-verb-object structure and preference modality. Similarly, the gloss *“IX SON FOOTBALL LEAGUE ADDRESS WHERE DO DO”* is rendered as *“Can you give me the address of the football league where your son attends?”*, preserving both WH-question formation and possessive relations. A more subtle example is *“TIRED WORK DONT WANT”*, which the model translates to *“I do not feel like going to work today, I am tired.”*, correctly resolving negation and causal conjunctions.

These examples reflect the model’s ability to handle syntactic reordering, implicit subjects, and complex questions. Over 85% of evaluated outputs were semantically aligned with references, validating the high BLEU-1 score (83.2) with strong real-world interpretability.

## 5.3 Data Augmentation

To improve generalization under diverse visual conditions, we designed a customized RandAugment-inspired data augmentation pipeline comprising both standard and enhanced transformations. We first apply random resized cropping to  $224 \times 224$  pixels, with scale sampled uniformly from  $[0.08, 1.0]$  and aspect ratio from  $[0.75, 1.33]$ , simulating partial visibility and varying image coverage. This is followed by color jittering with brightness, contrast, and saturation factors up to 0.4 and hue shifts up to 0.1. Horizontal flipping and color jittering are each applied with a probability of 0.5. These operations are implemented on-the-fly during training for efficiency. To model interactions between multiple distortion types and simulate complex visual variations, we introduce a multi-stage augmentation strategy. For every  $N_{th}$  frame in a video sequence, we apply  $M$  random transformations sampled from a set

$\mathcal{F}_{aug} = \{\text{scaling, rotation, affine, perspective, elastic}\}$ . Each transformation is constrained within a 50% parameter bound to preserve content integrity. Enhanced frames are integrated back into the original sequence to preserve temporal alignment, encouraging the model to track gesture features under increasingly diverse visual perturbations. This cascading augmentation strategy effectively exposes the network to a broader manifold of input conditions, helping it learn more robust and user-generalizable representations across ASL signing styles.

## 6 EVALUATION

The evaluation of the SignGlass system focuses on assessing SignGlass accuracy and overall effectiveness in recognizing ASL and translating it into coherent spoken English. This section details the implementations, results from user studies, and robustness of evaluation.

### 6.1 Implementation

Our SignGlass algorithms are implemented with Pytorch [59] library, and the training is performed on a server with Intel AMD EPYC 74F3 CPU, 128GB RAM, and four NVIDIA RTX A6000 GPU. We set the batch size to 2 and the learning rate to  $1e-4$ . We use the Adam optimizer. To avoid potential overfitting during the training process, we apply the L2 regularization with a parameter of 0.0001. Once the model is generated from training, the inference is performed on smartphones (Samsung S20 and OnePlus 9 Pro) using Pytorch Mobile [26].

### 6.2 User Study

**6.2.1 Dataset.** To validate our approach, we selected a representative set of ASL expressions that include both manual and non-manual markers. We used professional ASL data from the How2Sign dataset [24] and the ASLLRP Sign Bank [9], which span topics such as sports, entertainment, personal care, education, and home and garden. A total of 600 sentences were chosen, evenly distributed

across five non-manual categories: negative, WH-questions, Yes/No-questions, conditionals, and topicalization. This selection ensures broad coverage of sentence types.

**6.2.2 Data Collection.** Our study was approved by the Institutional Review Board (IRB). As shown in Fig. 7, participants wore the smart glasses described in Section 4.3. We recruited 14 Deaf participants (10 females and 4 males), all native ASL users, aged between 20 and 50 years. Fig. 11 illustrates our carefully designed user study setup conducted in a quiet conference room to ensure consistent environmental conditions.

While wearing SignGlass, participants watched ASL expression videos and were instructed to understand and re-sign corresponding ASL expressions. Each participant lasted approximately two hours, including three scheduled breaks during which the glasses were removed and remounted to simulate natural variations in usage and improve system robustness. In general, each participant signed more than 200 ASL expressions. This carefully structured data collection process allowed us to capture a wide range of manual and non-manual ASL features in realistic scenarios, laying a strong foundation for robust system development and evaluation.

During data collection, we used a camera with an FOV of 90° to record the ground truth of the manual marker and non-manual marker of the user. We built a local Ethernet network featuring high bandwidth connectivity between Raspberry Pi boards and the PC to ensure robust transmission and prevent frame dropping. The ground truth and those of Raspberry Pis are synchronized using timestamps. Annotators reviewed these recordings to check the alignment among ASL videos, gloss labels, and their English text.

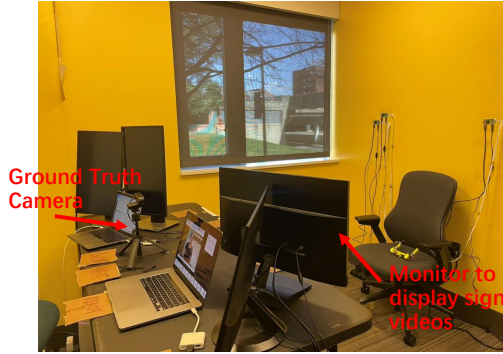


**Figure 11: Example of user study setups: users were seated in front of the monitor, and our team members were seated in the opposite direction.**

### 6.3 Evaluation Metrics

To show the effectiveness of ASL recognition and translation, we utilize the Word Error Rate (WER) to evaluate the recognition accuracy (input videos → ASL glosses), and the Bilingual Evaluation Understudy (BLEU) score to evaluate the translation quality (ASL glosses → English). Both ASL glosses and English are created and verified by ASL experts.

**Word Error Rate (WER).** WER is a critical metric for evaluating the accuracy of ASL recognition systems. It quantifies how effectively the system identifies ASL glosses from input video sequences.

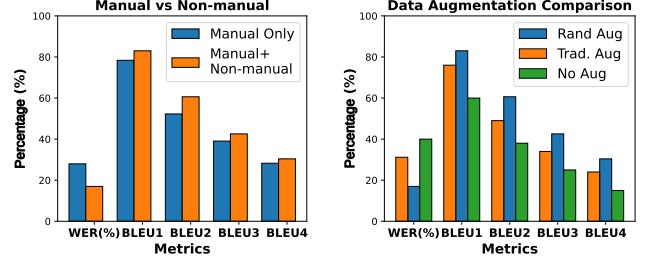


**Figure 12: Comparison of Manual vs. Multi-model with Different Metrics** **Figure 13: Comparison of Manual vs. Multi-model with vs. without Augmentation with Different Metrics**

Specifically, WER is computed by comparing the predicted gloss sequence to a reference sequence, which includes both manual markers (e.g., *BOOK*, *GIVE*) derived from hand gestures and non-manual markers (e.g., *WHQ*, *NEG*) originating from facial expressions or head movements. The WER is defined as:  $WER = \frac{S + I + D}{N}$ , where  $S$ ,  $I$ , and  $D$  denote the number of substitutions, insertions, and deletions, respectively, and  $N$  is the total number of words in the reference. A lower WER indicates higher recognition accuracy and overall system performance.

**Bilingual Evaluation Understudy (BLEU) Score.** The BLEU score is a widely adopted metric for evaluating machine translation quality by comparing a generated sentence to one or more reference translations [18, 58]. In our work, BLEU is used to assess the effectiveness of translating ASL gloss sequences into natural English text. The BLEU score is computed as:

$BLEU = \exp \left( \min \left( 0, 1 - \frac{r}{c} \right) + \sum_{n=1}^4 w_n \log p_n \right)$ , where  $r$  is the length of the reference translation,  $c$  is the length of the candidate (translated) sentence,  $w_n$  denotes the weight for each  $n$ -gram level, and  $p_n$  is the precision score for  $n$ -grams of size  $n$ . We report precision for BLEU-1 (unigram), BLEU-2 (bigram), BLEU-3 (trigram), and BLEU-4 (four-gram), which together reflect the translation's fluency and syntactic correctness at multiple linguistic levels. Higher BLEU scores indicate closer alignment between the generated and reference sentences, signifying higher translation quality.

### 6.4 Overall Performance

**6.4.1 Overall Performance on Unseen Users.** In this section, we present the overall performance evaluation using the leave-one-user-out method, where data from one previously unseen user is tested. Additionally, we will discuss the performance across various sentence categories to provide a detailed analysis of different non-manual categories.

**Leave-one-user-out Performance.** Figure 14 illustrates the WER and BLEU scores across all 14 users, comparing the performance of manual-only recognition against our proposed multi-modal approach. We evaluate the system using leave-one-user-out cross-validation over 14 users, where in each iteration, the model is trained on data from 13 users and tested on the remaining one. This process is repeated 14 times to ensure that each user's data is used for evaluation exactly once. As shown in Fig.12, the WER for

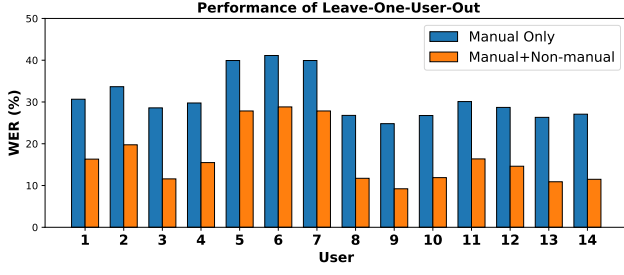


Figure 14: Comparison of Manual vs. Multi-model for Leave-one-out with Different Users

ASL recognition is 16.98%, which demonstrates the effectiveness of recognizing the manual and non-manual markers. The BLEU1 score is 83%, which proves that our system is effective in translating the comprehensive glosses into English. Notably, the incorporation of non-manual markers led to a substantial reduction in the WER score from 31.17% to 16.98% and led to an increase in BLEU1, BLEU2, BLEU3, and BLEU4 scores from 68.8% to 83.0%, from 47.3% to 68.9%, from 32.5% to 57.3%, and from 22.5% to 47.4%, respectively. The dramatic reduction in WER proves that our innovative system is successfully recognizing and translating the complex interplay between manual markers and non-manual markers (e.g., subtle facial expressions and body postures). In addition, the consistent improvements from BLEU1 through BLEU4 indicate that our system maintains its effectiveness even as linguistic complexity increases. These results represent a significant achievement in ASL recognition and translation.

**Compared with State-of-the-Art ASL Recognition Systems:** Compared with the other two wearable systems supporting non-manual markers recognition, SmartASL [41] and TransASL[37], our system has two advantages: 1) only use a pair of smart glasses instead of multiple devices to sense manual and non-manual marker recognition, and the earbuds are always not socially acceptable in the deaf community because the earbuds are not useful for deaf people, and 2) they can only achieve acceptable user-dependent performance with 40 sentences during controlled environments. In contrast, our system only used a pair of glasses, which is more acceptable for the deaf community, and achieved user-independent performance (WER=16.98%, BLEU1=83%), which indicates no extra sample data collection and training efforts. Thus, our system can be deployed to any new user directly, thus improving the generalizability in daily life. In summary, our system, SignGlass, represents a significant advancement by offering a user-friendly, affordable, and accessible technology designed to bridge the communication gap between sign language users and spoken language users. This innovative solution paves the way for enhanced communication and understanding, making it an invaluable tool for improving accessibility.

**Performance Across Various Sentence Categories.** Given that non-manual markers are crucial for forming various categories of English sentences, including negatives, Wh-questions, yes/no questions, condition (C/W), and topics [53], we have categorized all sentences into six groups. As depicted in Table.12, the Word Error Rate (WER) varies from 8.35% to 27.41%, indicating promising

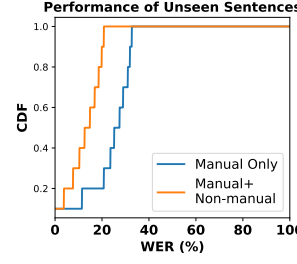


Figure 15: Performance of Unseen Sentence

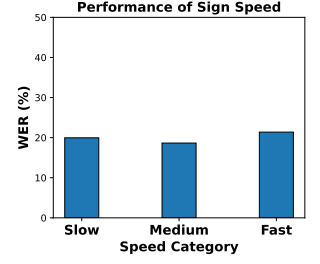


Figure 16: Performance of Different Sign Speed.

performance across different ASL sentence categories. Notably, the yes/no question category exhibited the lowest WER at 8.35%. This may be attributed to the distinctive yes/no question non-manual markers, especially for head shakes, which significantly alter the visual presentation in two-sided camera views. In contrast, the C/W category has the highest WER, which means the wrong classification, which may be attributed to the no-obvious non-manual markers, such as eyebrows. Thus, it is necessary to explore the potential solutions for fine-grained non-manual marker recognition.

## 6.5 Overall Performance On Augmentation Enhancement

While traditional data augmentation methods have been shown to enhance data diversity and align with user preferences, they have not demonstrated the same level of effectiveness in our specific application. Thus, we proposed an advanced algorithm to effectively enhance the diversity of different manual and non-manual markers thoroughly. In this section, we will compare these two data augmentation methods: baseline to represent the traditional data augmentation method and enhancement to represent our proposed data augmentation method. As illustrated in Fig.13, we achieved approximately a 10% reduction in WER, highlighting a considerable improvement in ASL recognition accuracy. These results collectively showcase the effectiveness of our augmentation method. By incorporating more nuanced variations and capturing subtle features, we've enhanced model accuracy while also notably reducing computational demands.

## 6.6 Statistic Performance

**6.6.1 Impact of Unseen Sentences.** In real-world deployment, ASL translation systems must handle sentences that differ from those seen during training, due to the inherent variety in sign vocabulary, grammar, and phrasing. To evaluate SignGlass's robustness on unseen sentences, we construct an evaluation set in which sentence-level entries are held out from training. Although the individual signs may still appear in the training data, these test sequences exhibit novel combinations and syntactic structures.

As shown in Fig. 15, the cumulative distribution function (CDF) of WER reveals that SignGlass with both manual and non-manual channels consistently outperforms the manual-only baseline. Over 80% of the unseen test sentences have WER below 25% when using our full model, while the manual-only model exhibits significantly

| Sentence Categories | WER    | BLEU-1 Score | BLEU-2 Score | BLEU-3 Score | BLEU-4 Score |
|---------------------|--------|--------------|--------------|--------------|--------------|
| NEG                 | 18.33% | 81.67%       | 66.70%       | 54.50%       | 44.50%       |
| WH-Q                | 15.68% | 84.32%       | 71.10%       | 59.90%       | 50.40%       |
| Y/N                 | 8.35%  | 91.65%       | 84.00%       | 77.00%       | 70.60%       |
| C/W                 | 27.41% | 72.59%       | 52.70%       | 38.40%       | 27.90%       |
| TOP                 | 17.35% | 82.65%       | 68.30%       | 56.50%       | 46.70%       |
| Others              | 17.71% | 82.29%       | 67.70%       | 55.50%       | 45.50%       |

Table 3: Recognition and Translation Performance based on Different Sentence Categories

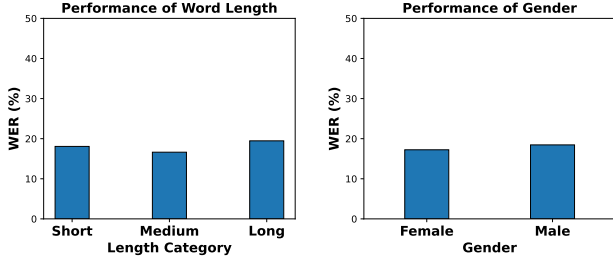


Figure 17: Performance of Different Gloss Lengths.

worse error rates. This indicates that our multi-modal architecture generalizes better to new compositions of signs, owing to its capacity to model contextual dependencies and facial grammar cues.

**6.6.2 Impact of ASL Sign Speed.** Signers naturally differ in fluency [51], leading to significant variation in the speed at which ASL is performed. This variation can pose challenges for recognition systems, particularly when dealing with fast, fluid gestures that compress temporal features or introduce motion blur.

To evaluate SignGlass’s robustness to varying sign speeds, we compute the average number of signs performed per second (*sign speed*) for each video, defined as the length of the ASL gloss divided by the corresponding video duration. Based on this continuous metric, we categorize the videos into three speed bins: **Slow** (0–2 signs/sec), **Medium** (2–4 signs/sec), and **Fast** (4–6 signs/sec). This categorization reflects realistic fluency ranges observed in our participant cohort.

As shown in Fig. 16, the Word Error Rate (WER) generally increases with sign speed. SignGlass achieves its lowest WER in the Medium group (11.55%), suggesting an optimal balance between motion clarity and temporal density. In contrast, the WER rises to 15.79% in the Fast category and slightly drops to 12.89% in the Slow group, although this fluctuation is less pronounced. The performance degradation at high speeds likely results from reduced spatial clarity and faster transitions between signs, which challenge the system’s temporal modeling capabilities.

These findings demonstrate SignGlass’s ability to handle a broad range of signing speeds, with moderate degradation under rapid signing conditions. This robustness is critical for real-world deployment, where users may exhibit diverse levels of fluency and tempo.

**6.6.3 Impact of ASL Gloss Length .** The length of ASL glosses within a sentence can significantly affect translation performance,

as it determines the amount of contextual information available to the model. To assess SignGlass’s robustness to sequence length, we categorize glosses into three groups based on the number of signs per sentence: **Short** (1–4 signs), **Medium** (5–8 signs), and **Long** (9–12 signs).

As shown in Fig. 17, SignGlass demonstrates improved performance on longer glosses, achieving the lowest Word Error Rate (WER) of 1.90% in the Long category. In contrast, the Short category yields the highest WER at 25.83%. This result aligns with the characteristics of the underlying temporal encoder (LSTM), which benefits from richer context in longer sequences, allowing for more accurate modeling of inter-sign dependencies. On the other hand, short sequences offer less temporal information and are more prone to recognition ambiguity due to the absence of contextual cues.

Despite this variation, SignGlass maintains acceptable performance across all length groups, indicating its adaptability to real-world usage where utterances vary widely in complexity and length.

**6.6.4 Impact on Gender.** Physiological differences between individuals—such as hand size, facial shape, and movement styles—can potentially influence the accuracy of sign language recognition systems. To evaluate whether SignGlass is sensitive to such differences, we analyze performance across two user groups based on physiological gender, as collected during the participant enrollment phase.

Fig. 18 reports the Word Error Rate (WER) for male and female signers. The average WER for female participants is 16.17%, while male participants exhibit an average of 18.68%. Although some individual variation exists within each group, especially at the higher end of the range, the overall performance difference is not statistically significant.

These results suggest that SignGlass provides consistent recognition accuracy across different gender groups, supporting its suitability for inclusive, real-world deployment across diverse user populations.

**6.6.5 Impact of Quantity of Cameras.** SignGlass employs a three-camera configuration to capture both manual and non-manual sign components. To evaluate how different camera configurations affect system performance, we compare three setups: **One Camera** (hand view only), **Two Cameras** (one side face + hand), and **Three Cameras** (left face + right face + hand).

As shown in Fig. 19, recognition accuracy improves with the number of cameras. The one-camera setup yields the highest Word Error Rate (WER) and lowest BLEU scores, indicating limited modality coverage. Adding a second camera significantly enhances performance by introducing non-manual cues from the face. The full three-camera configuration achieves the lowest WER and highest

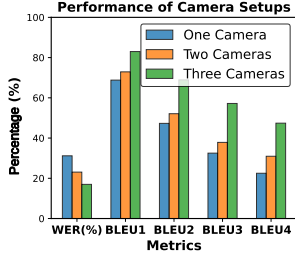


**Figure 19: Performance of Different Camera Numbers.**

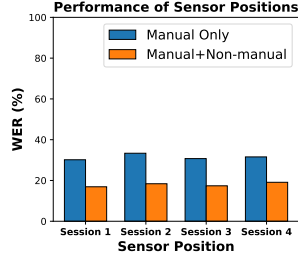


**Figure 20: Performance of Different Sensor Positions.**

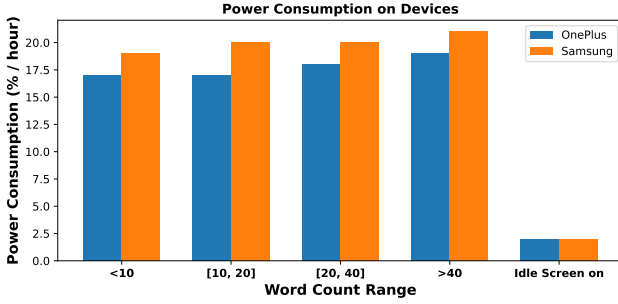


**Figure 21: Power consumption of executing ML models in SignGlass on mobile device**

BLEU scores, confirming that comprehensive spatial coverage is essential for robust sign language understanding.

These results underscore the importance of multimodal input from multiple viewpoints, particularly for capturing subtle facial expressions and hand-face interactions. While fewer cameras reduce hardware complexity, the performance trade-offs suggest that three viewpoints are necessary for high-fidelity ASL translation.

**6.6.6 Impact of Sensor Position.** Wearable systems are subject to inevitable variability in how users wear the device, including slight changes in tilt, height, or alignment. To test the robustness of SignGlass under such conditions, we conducted a user study across four independent wearing sessions. In each session, users removed and remounted the wearable device using adjustable nose pads, resulting in small variations in camera and sensor positions.

Fig. 20 shows the system’s WER across the four sessions. Both the manual-only and the combined manual + non-manual configurations maintain stable performance across sessions, with no significant degradation caused by repositioning. We attribute this consistency to the snug and repeatable fit of the device’s mounting system, which limits the range of positional deviation.

These results suggest that SignGlass is robust to normal re-wearing scenarios and does not require precise alignment for each use, making it practical for daily deployment in real-world settings.

**6.6.7 Power Consumption.** For wearable ASL translation systems to be viable in daily use, power efficiency is critical—especially when running inference locally on mobile devices. To assess the energy impact of SignGlass, we profile the power consumption of

its machine learning pipeline running on two commercial smartphones: the Samsung S20 and the OnePlus 9 Pro. The measurement is conducted using Android Battery Stats and Battery Historian, focusing solely on inference workloads and excluding video capture or preprocessing stages.

As shown in Fig. 21, SignGlass maintains consistent power consumption across both devices. During active inference, the OnePlus 9 Pro exhibits a power discharge rate of 17% per hour, while the Samsung S20 consumes slightly more at 20% per hour. These values are well within acceptable limits for continuous use over several hours, affirming the feasibility of edge deployment.

This level of efficiency enables SignGlass to operate in real-time on off-the-shelf mobile hardware without relying on external computing infrastructure. The results demonstrate that our system supports sustainable, on-device ASL translation suitable for wearable, everyday applications.

**6.6.8 Ablation Study of Core Modules.** To validate the contribution of each individual component in our SignGlass pipeline, we conducted an ablation study by selectively removing or replacing key modules. All models were evaluated using the leave-one-user-out protocol across the same dataset, ensuring comparability. Table 4 summarizes the Word Error Rate (WER) and BLEU1 score for each variant.

**Table 4: Ablation Results of Key Modules**

| Model Variant                    | WER (%)      | BLEU1 (%)   |
|----------------------------------|--------------|-------------|
| <b>SignGlass (Ours)</b>          | <b>16.98</b> | <b>83.0</b> |
| w/o EgoCorrNet                   | 21.3         | 75.4        |
| w/o Optical Flow (landmark only) | 42.9         | 52.9        |
| w/o Landmark Input (flow only)   | 68.5         | 24.4        |
| w/o Cross-Modal Alignment        | 24.7         | 73.0        |
| w/o Dual TimeSFormer             | 24.1         | 72.5        |
| w/o FiLM Conditioning            | 19.7         | 78.8        |

**EgoCorrNet.** Replacing our EgoCorrNet module with a standard Bi-LSTM encoder led to a 4.3% increase in WER, indicating the effectiveness of spatio-temporal correlation in modeling egocentric hand movements under jitter and partial occlusion.

**Optical Flow Ablation.** Removing the optical flow input and retaining only hand landmark heatmaps leads to a sharp performance drop—WER rises to 42.9%, and BLEU1 drops to 52.9%. This indicates that landmark features alone are insufficient to capture temporal dynamics in continuous signing, especially under rapid motion or occlusion. Optical flow provides dense, motion-aware cues that complement the sparse skeletal structure, enabling better tracking of gesture transitions.

**Landmark Ablation.** Removing the landmark stream and using only optical flow results in catastrophic degradation, with WER rising to 68.5% and BLEU1 falling to just 24.4%. Although flow captures coarse motion, it lacks semantic grounding and fine spatial detail—particularly in distinguishing finger articulations and localized handshapes. These results confirm that flow and landmark cues are not redundant, but instead form a necessary dual-modality input for robust recognition.

**Jitter-Aware Alignment Ablation.** Disabling the jitter-aware cross-modal alignment module and directly fusing optical flow and landmark features without temporal attention leads to a 7.7% WER increase and a 10-point BLEU1 drop. This suggests that passive fusion cannot compensate for misaligned motion signals caused by egocentric jitter or self-occlusion. The attention-based alignment allows each modality to adaptively correct for temporal inconsistencies in the other, leading to more stable gesture embeddings.

**Dual TimeSFormer.** Using only a single facial view significantly degraded performance, especially in cases where symmetrical expressions (e.g., eyebrow raises for Yes/No questions) were required. BLEU1 dropped by over 10%, confirming the necessity of dual-view facial modeling.

**FiLM Conditioning.** Removing FiLM conditioning from the facial pathway reduced the model’s ability to generalize to signers with varied facial expressions, resulting in a 4.2% BLEU1 drop.

These results validate the architectural choices in SignGlass. Each module plays a critical role in enhancing robustness, cross-user generalizability, and overall ASL-to-English translation quality.

## 7 DISCUSSION

### 7.1 Potential Privacy Issue Discussion.

A significant concern with traditional front camera-based ASL recognition and translation systems is the potential privacy risk posed by the continuous recording of the user’s face, which could be misused. These recordings may inadvertently capture sensitive or private moments, raising privacy concerns. In contrast, as illustrated in Fig.7, our approach records only the side view of the user’s face, thereby minimizing the capture of full facial information and reducing the risk of privacy breaches. Moreover, our system is designed to solely record data pertaining to the system users, coupled with the implementation of stringent permission management protocols, including face authentication, to safeguard against unauthorized use of the recordings. This mechanism also serves to address privacy concerns related to individuals interacting with the signing user. Compared to other ASL sensing technologies, such as acoustic sensing [37, 40] and mmWave sensing [64], our system offers enhanced privacy protection while maintaining high performance by capturing all essential ASL components.

### 7.2 Potential Two-way Communication

Our current design of SignGlass assists DHH individuals by translating ASL into spoken English, thereby facilitating communication between DHH individuals and hearing people. However, SignGlass can not assist in understanding the hearing people for DHH individuals, limiting the user experience of conversation, because additional devices such as smartphones are often necessary to translate the speech for DHH individuals. Thus, we aim to integrate voice and text input for DHH individuals, with transcriptions displayed on the smart glasses. To further improve the user experience, future versions will also explore predictive responses and more seamless ASL–English–ASL conversations to foster inclusive communication.

### 7.3 Translation Errors Check

Given the potential for occasional mispredictions in ASL recognition, it is crucial to develop an advanced solution that enables ASL

signers to correct these errors effectively. One promising approach involves displaying the recognized ASL text transcription on smart glasses to provide immediate feedback. Besides, recent advances in LLMs have demonstrated substantial capability in detecting logical coherence within conversations, specifically through identifying semantic consistency, logical cause-and-effect relationships, and clear referents across dialogue. Leveraging these abilities, LLMs could potentially be employed to identify and flag logical errors in dialogue that may arise during communication, thereby improving the reliability of ASL interpretation.

### 7.4 Potential Solutions for Inter-user Variability

It is widely acknowledged that inter-user variability in ASL signing exists. To address this variability and further improve the ASL interpreting accuracy, it is required to explore potential solutions. Firstly, additional sensing modalities, such as IMU sensors on smartwatches or rings, will be incorporated to capture fine-grained motion dynamics, providing complementary temporal and kinematic cues to better distinguish gesture variations arising from differences in speed, articulation, or physical style across users. In addition, we plan to explore dynamic time warping (DTW) as a preprocessing step to temporally align gestures performed at different paces, improving model robustness to signing speed discrepancies. Furthermore, with the growing capabilities of LLMs, we are investigating their use for semantic inference and ambiguity resolution, enabling the system to generalize across diverse signing styles and gesture expressions. Together, these enhancements aim to build a more adaptive and context-aware system that effectively learns from and accommodates inter-user variability.

### 7.5 Form Factor Optimization and Future Development of ASL Smartglasses

So far, SignGlass demonstrates the feasibility of recognizing ASL sentences using smart glasses; however, further development is needed to transition it into a deployable product. The current prototype leverages off-the-shelf electronics for rapid development and validation, but its form factor remains bulky and suboptimal for long-term wear. In future iterations, we plan to develop a custom PCB with compact microcontrollers (e.g., ESP32-S3) and camera multiplexer ICs, enabling a more compact, fully integrated smart glasses system without relying on an external compute puck. We also aim to implement local inference on the glasses to support on-device processing, although limited hardware may affect model accuracy and performance. We acknowledge that alternative form factors, such as smartwatches or earphones, offer better social acceptability and comfort for daily use. However, these modalities typically capture only partial gestures, limiting full-sentence ASL translation. Future work will focus on balancing form factor, wearability, and expressive capability to ensure usability and comprehensive gesture understanding.

### 7.6 Potential AI Interactions

To enable seamless and natural ASL-based interactions with conversational agents, we plan to integrate our system with AI chatbots and large language models (LLMs), enhancing digital accessibility for DHH users. By leveraging the expressive richness of ASL and the contextual understanding of state-of-the-art LLMs, our system

will allow users to engage in fluid, real-time conversations with AI-driven assistants using signed input instead of typed or spoken language. This integration facilitates intuitive communication and bridges critical accessibility gaps in current conversational AI platforms, which often rely solely on voice or text. Through this approach, we aim to support diverse use cases—from daily queries and task management to emotional support and educational assistance—tailored to the needs and preferences of the DHH community. Ultimately, SignGlass advances the broader goal of inclusive human-AI interaction by making conversational technologies more responsive, equitable, and usable for sign language users.

## 8 CONCLUSION

In this study, we introduced SignGlass, an end-to-end smartglass-based ASL recognition and translation system designed to detect and interpret real ASL sentences by integrating manual and non-manual markers. SignGlass leverages a downward-facing wearable camera positioned in the center of the glasses to capture manual markers. In contrast, two inward-facing cameras on the sides of the glasses capture non-manual markers. We evaluated the system using data collected from 14 Deaf users in a user-independent training setup. The results demonstrated that SignGlass can translate ASL into English with a Word Error Rate (WER) of 16.98% and a BLEU score of 83%. These promising results suggest that SignGlass has significant potential as a practical and ubiquitous system to facilitate daily communication between ASL users and hearing individuals.

## ACKNOWLEDGMENTS

We are grateful to all anonymous reviewers for their constructive comments. This work is supported in part by the NSF grants: CAREER IIS-2046972.

## REFERENCES

- [1] 2017. CyberGlove. <http://www.cyberglovesystems.com/>.
- [2] 2017. OV5647 Mini Camera. <https://www.arducam.com/product/b006605-arducam-for-raspberry-pi-zero-camera-module-wide-angle-160-1-4-inch-5mp-ov5647-spy-camera-with-flex-cable-for-pi-zero-and-pi-compute-module/>.
- [3] 2021. TapStrap 2. <https://www.tapwithus.com/>.
- [4] 2023. ASL Sign for BIKE. <https://www.lifeprint.com/asl101/pages-signs/b/bicycle.htm>.
- [5] 2023. ASL Sign for EAT. <http://www.lifeprint.com/asl101/pages-signs/e/eat.htm>.
- [6] 2023. World Federation of the Deaf. <https://wfdeaf.org/our-work/>.
- [7] Amit Adate, Soroush Shahi, Rawan Alharbi, Sougata Sen, Yang Gao, Aggelos K Katsaggelos, and Nabil Alshurafa. 2022. Detecting screen presence with activity-oriented RGB camera in egocentric videos. In *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. IEEE, 403–408.
- [8] Oya Aran. 2008. Vision based sign language recognition: modeling and recognizing isolated signs with manual and non-manual components. *Bogaziçi University* (2008).
- [9] ASLLRP. [n. d.]. ASLLRP ASL Sign Bank. <https://asllrp.caset.buffalo.edu/signbank.html>. Accessed: 2023-11-01.
- [10] Benjamin J Bahan. 1997. Non-manual realization of agreement in American Sign Language. (1997).
- [11] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is Space-Time Attention All You Need for Video Understanding?. In *International Conference on Machine Learning (ICML)*.
- [12] Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreaux, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoeft, et al. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 16–31.
- [13] Necati Cihan Camgoz et al. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *CVPR*.
- [14] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7784–7793.
- [15] George Caridakis, Stylianos Asteriadis, and Kostas Karpouzis. 2014. Non-manual cues in automatic sign language recognition. *Personal and ubiquitous computing* 18, 1 (2014), 37–46.
- [16] Tuochao Chen, Yaxuan Li, Songyun Tao, Hyunchul Lim, Mose Sakashita, Ruidong Zhang, François Guimbretière, and Cheng Zhang. 2021. Neckface: Continuously tracking full facial expressions on neck-mounted wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–31.
- [17] Tuochao Chen, Benjamin Steeper, Kinan Alsheikh, Songyun Tao, François Guimbretière, and Cheng Zhang. 2020. C-face: Continuously reconstructing facial expressions by deep learning contours of the face with ear-mounted miniature cameras. In *Proceedings of the 33rd annual ACM symposium on user interface software and technology*. 112–125.
- [18] Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. 2022. Two-stream network for sign language recognition and translation. *Advances in Neural Information Processing Systems* 35 (2022), 17043–17056.
- [19] Seokmin Choi, Yang Gao, Yincheng Jin, Se jun Kim, Jiyang Li, Wenyao Xu, and Zhanpeng Jin. 2022. PPGface: Like What You Are Watching? Earphones Can" Feel" Your Facial Expressions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–32.
- [20] Necati Cihan Camgoz et al. 2017. Subnets: End-to-end hand shape and continuous sign language recognition. In *ICCV*.
- [21] Deaf Community. 2021. Deaf Culture. <https://www.startasl.com/what-does-d-d-and-d-deaf-mean-in-the-deaf-community/>. [Updated May 13, 2021].
- [22] ASLLRP Continuous Signing Corpora. 2022. American Sign Language Linguistic Research Project. <https://dai.cs.rutgers.edu/dai/s/dai>. [Online].
- [23] Cao Dong, Ming C Leu, and Zhaozheng Yin. 2015. American sign language alphabet recognition using microsoft kinect. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 44–52.
- [24] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metz, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2sign: a large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2735–2744.
- [25] Deniz Ekiz, Gamze Ege Kaya, Serkan Buğur, Sila Güler, Buse Buz, Bilgin Kosucu, and Bert Arnrich. 2017. Sign sentence recognition with smart watches. In *2017 25th Signal Processing and Communications Applications Conference (SIU)*. IEEE, 1–4.
- [26] Facebook. 2021. Introduce lite interpreter workflow in Android and iOS. "https://pytorch.org/mobile/android/".
- [27] Meta AI Glass. 2022. Meta AI Glass. <https://www.meta.com/ai-glasses/>.
- [28] Google. 2022. AR Glass. <https://nerdist.com/article/google-ar-glasses-live-translation-real-time-transcription/>.
- [29] Yunqi Guo et al. 2023. Sign-to-911: Emergency Call Service for Sign Language Users with Assistive AR Glasses. In *MobiCom*.
- [30] Aiming Hao et al. 2021. Self-mutual distillation learning for continuous sign language recognition. In *CVPR*.
- [31] Jiahui Hou, Xiang-Yang Li, Peide Zhu, Zefan Wang, Yu Wang, Jianwei Qian, and Panlong Yang. 2019. SignSpeaker: A real-time, high-precision smartwatch-based sign language translator. In *The 25th Annual International Conference on Mobile Computing and Networking (MobiCom'19)*. Article 24, 15 pages.
- [32] Hezhen Hu et al. 2021. Hand-model-aware sign language recognition. In *AAAI*.
- [33] Hezhen Hu et al. 2021. Signbert: pre-training of hand-model-aware representation for sign language recognition. In *ICCV*.
- [34] Lianyu Hu, Wei Feng, Liqing Gao, Zekang Liu, and Liang Wan. 2024. CorrNet+: Sign Language Recognition and Translation via Spatial-Temporal Correlation. *arXiv preprint arXiv:2404.11111* (2024).
- [35] Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. 2018. Video-based sign language recognition without temporal segmentation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- [36] Google Inc. 2022-05-15. Google teases smart glasses prototype that translates languages in real time. <https://www.cnn.com/2022/05/11/google-smart-glasses-prototype-translates-languages-in-real-time.html>.
- [37] Yincheng Jin, Seokmin Choi, Yang Gao, Jiyang Li, Zhengxiong Li, and Zhanpeng Jin. 2023. TransASL: A Smart Glass based Comprehensive ASL Recognizer in Daily Life. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 802–818.
- [38] Yincheng Jin, Yang Gao, Xiaotao Guo, Jun Wen, Zhengxiong Li, and Zhanpeng Jin. 2022. EarHealth: an earphone-based acoustic otoscope for detection of multiple ear diseases in daily life. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*. 397–408.
- [39] Yincheng Jin, Yang Gao, Xuhai Xu, Seokmin Choi, Jiyang Li, Feng Liu, Zhengxiong Li, and Zhanpeng Jin. 2022. EarCommand: "Hearing" Your Silent Speech

- Commands In Ear. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–28.
- [40] Yincheng Jin, Yang Gao, Yanjun Zhu, Wei Wang, Jiyang Li, Seokmin Choi, Zhangyu Li, Jagmohan Chauhan, Anind K Dey, and Zhanpeng Jin. 2021. SonicASL: An acoustic-based sign language gesture recognizer using earphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–30.
- [41] Yincheng Jin, Shibo Zhang, Yang Gao, Xuhai Xu, Seokmin Choi, Zhengxiong Li, Henry J Adler, and Zhanpeng Jin. 2023. SmartASL: "Point-of-Care" Comprehensive ASL Interpreter Using Wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 2 (2023), 1–21.
- [42] Sara Askari Khomami and Sina Shamekhi. 2021. Persian sign language recognition using IMU and surface EMG sensors. *Measurement* 168 (2021), 108471.
- [43] Oscar Koller, Jens Forster, and Hermann Ney. 2017. Weakly supervised learning with multi-instance dynamic time warping and hidden Markov models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4775–4784.
- [44] Jiyang Li et al. 2023. SignRing: Continuous American Sign Language Recognition Using IMU Rings and Virtual IMU Data. *ACM IMWUT* (2023).
- [45] Jiyang Li, Lin Huang, Siddharth Shah, Sean J Jones, Yincheng Jin, Dingran Wang, Adam Russell, Seokmin Choi, Yang Gao, Junsong Yuan, et al. 2023. SignRing: Continuous American Sign Language Recognition Using IMU Rings and Virtual IMU Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–29.
- [46] Hyunchul Lim, Nam Anh Dang, Dylan Lee, Tianhong Catherine Yu, Jane Lu, Franklin Mingzhe Li, Yiqi Jin, Yan Ma, Xiaojun Bi, François Guimbretière, et al. 2025. SpellRing: Recognizing Continuous Fingerspelling in American Sign Language using a Ring. *arXiv preprint arXiv:2502.10830* (2025).
- [47] J Liu, B Liu, S Zhang, F Yang, P Yang ... Gesture Recognition ..., and undefined 2013. 2013. Recognizing eyebrow and periodic head gestures using CRFs for non-manual grammatical marker detection in ASL. *ieeexplore.ieee.org* (2013). <https://ieeexplore.ieee.org/abstract/document/6553781/> Employed a 2-level CRF framework for ASL recognition, focusing on eyebrow and head gestures' temporal phases using geometric and appearance features..
- [48] Yilin Liu et al. 2020. Finger Gesture Tracking for Interactive Applications: A Pilot Study with Sign Languages. *ACM IMWUT* (2020).
- [49] Yilin Liu, Fengyang Jiang, and Mahanth Gowda. 2020. Finger Gesture Tracking for Interactive Applications: A Pilot Study with Sign Languages. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–21.
- [50] Denys JC Matthies, Bernhard A Strecker, and Bodo Urban. 2017. Earfieldsensing: A novel in-ear electric field sensing to enrich wearable gesture input through facial expressions. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 1911–1922.
- [51] Rachel I Mayberry and Robert Kluender. 2018. Rethinking the critical period for language: New insights into an old question from American Sign Language. *Bilingualism: Language and Cognition* 21, 5 (2018), 886–905.
- [52] Dimitris Metaxas, Bo Liu, Fei Yang, Peng Yang, Nicholas Michael, and Carol Neidle. 2012. Recognition of Nonmanual Markers in American Sign Language (ASL) Using Non-Parametric Adaptive 2D-3D Face Tracking. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey, 2414–2420.
- [53] Dimitris N Metaxas, Bo Liu, Fei Yang, Peng Yang, Nicholas Michael, and Carol Neidle. 2012. Recognition of Nonmanual Markers in American Sign Language (ASL) Using Non-Parametric Adaptive 2D-3D Face Tracking.. In *LREC*. 2414–2420.
- [54] Nicholas Michael, Peng Yang, Qingshan Liu, Dimitris N Metaxas, Carol Neidle, and CBIM Center. 2011. A Framework for the Recognition of Nonmanual Markers in Segmented Sequences of American Sign Language.. In *BMVC*. 1–12.
- [55] Yuecong Min et al. 2021. Visual alignment constraint for continuous sign language recognition. In *CVPR*.
- [56] Mohamed Mohandes, Mohamed Deriche, and Junzhao Liu. 2014. Image-based and sensor-based approaches to Arabic sign language recognition. *IEEE Transactions on Human-Machine Systems* 44, 4 (2014), 551–557.
- [57] United Nations. 2022. International day of sign language. <https://www.un.org/en/observances/sign-languages-day>. [Online].
- [58] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.
- [59] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019), 8026–8037.
- [60] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. 2018. FiLM: Visual Reasoning with a General Conditioning Layer. In *AAAI Conference on Artificial Intelligence*.
- [61] Deborah Chen Pichler. 2002. Word order variation and acquisition in American Sign Language. (2002).
- [62] Megan Salisbury and Jack Roberts. 2020. Wearables for deaf users: Challenges with ear-based devices. In *Proceedings of the 2020 ACM Conference on Human Factors in Computing Systems*. 752–765. <https://doi.org/10.1145/3342195.3387526>
- [63] Wendy Sandler. 2006. Sign language and linguistic universals. *Cambridge University* (2006).
- [64] Panneer Selvam Santhalingam, Al Amin Hosain, Ding Zhang, Parth Pathak, Huzefa Rangwala, and Raja Kushalnagar. 2020. mmASL: Environment-Independent ASL Gesture Recognition Using 60 GHz Millimeter-wave Signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1, Article 26 (2020), 30 pages.
- [65] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. Progressive transformers for end-to-end sign language production. In *European Conference on Computer Vision (ECCV)*. 687–705.
- [66] Thad Starner and Alex Pentland. 1997. Real-time american sign language recognition from video using hidden markov models. In *Motion-based recognition*. Springer, 227–243.
- [67] Karush Suri and Rinki Gupta. 2019. Continuous sign language recognition from wearable IMUs using deep capsule networks and game theory. *Computers & Electrical Engineering* 78 (2019), 493–503.
- [68] Noeru Suzuki, Yuki Watanabe, and Atsushi Nakazawa. 2020. Gan-based style transformation to improve gesture-recognition accuracy. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 4, 4 (2020), 1–20.
- [69] Clayton Valli. 2000. *Linguistics of American Sign Language: An Introduction*. Gallaudet University Press.
- [70] Dhruv Verma, Sejal Bhalla, Dhruv Sahnain, Jainendra Shukla, and Aman Parnami. 2021. ExpressEar: Sensing Fine-Grained Facial Expressions with Earables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–28.
- [71] Zhibo Wang et al. 2020. Hear sign language: A real-time end-to-end sign language recognition system. *IEEE Transactions on Mobile Computing* (2020).
- [72] Katharine L Watson. 2010. *WH-questions in American Sign Language: Contributions of non-manual marking to structure and meaning*. Purdue University.
- [73] Traci Patricia Weast. 2008. *Questions in American Sign Language: A quantitative analysis of raised and lowered eyebrows*. The University of Texas at Arlington.
- [74] WHO. 2022. Deafness and hearing loss. <https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss>. [Online].
- [75] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. 2020. MediaPipe Hands: On-device Real-time Hand Tracking. *arXiv:2006.10214 [cs.CV]* <https://arxiv.org/abs/2006.10214>
- [76] Qian Zhang, JiaZhen Jing, Dong Wang, and Run Zhao. 2022. WearSign: Pushing the Limit of Sign Language Translation Using Inertial and EMG Wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–27.
- [77] Qian Zhang, Dong Wang, Run Zhao, and Yinggang Yu. 2019. MyoSign: enabling end-to-end sign language recognition with wearables. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 650–660.
- [78] Siwei Zhao et al. 2023. A Cloud-Based Sign Language Translation System via CNN with Smart Glasses. In *International Conference on Green, Pervasive, and Cloud Computing*.
- [79] Hao Zhou et al. 2023. SignQuery: A Natural User Interface and Search Engine for Sign Languages with Wearable Sensors. In *MobiCom 2023*.
- [80] Hao Zhou et al. 2024. ASLRing: American Sign Language Recognition with Meta-Learning on Wearables. *IEEE/ACM IoTDI* (2024).
- [81] Ronglai Zuo and Brian Mak. 2022. C2slr: Consistency-enhanced continuous sign language recognition. In *CVPR*.