Comparing Sign Language AI and Human Interpreters: Evaluating Efficiency for Greater Social Sustainability

NICHOLAS WILKINS, Sign-Speak, USA
ZHIWEI YU, Sign-Speak, USA
NIKOLAS KELLY, Sign-Speak, USA
DENNIS MURPHY, Sign-Speak, USA
YAMILLET PAYANO, Sign-Speak, USA

While interpreters provide fantastic accessibility, equitable access in the Deaf and Hard of hearing (D/HH) community is deteriorating due to resource constraints. Not only are there rising demands being placed on the interpreting infrastructure, but additionally, there is a falling supply of qualified interpreters. Recent advances in automated sign language recognition and sign language avatar technology (AI sign language systems) have fostered opportunities for alleviating these constraints for D/HH people around the world. However, limited research has been done to test the deployment and efficiency of such innovations in practice. When human factors are considered, investigations are often superficial and fail to perform meaningful analysis against an understood baseline (such as interpreters). To the authors' knowledge, within the HCI body of research, no literature currently exists comparing the efficiency of these automated sign language systems and interpreters. Our work is a first step in pioneering such relevant analysis. This paper examines the efficiency of Sign Language Systems when compared to existing human interpreting infrastructures through the analysis of several experiments conducted with Deaf participants. In particular, we aim to establish baselines in accuracy and efficiency and formulate methodologies to analyze realistic deployment of AI tools to everyday signers through the contrasting of AI tools and human interpreters. Our results indicate that AI-powered sign language systems are non-inferior to human interpreters in terms of accuracy, efficiency, and robustness. These findings underscore the potential for AI-powered sign language systems to augment existing interpreting services, contributing to greater accessibility and autonomy for the D/HH community. We conclude with a brief discussion of the potential implications of integrating such technologies into daily life, including opportunities for enhanced communication and challenges that remain in ensuring equitable and ethical deployment.

1 Introduction

Across the globe, about 466 Million Deaf and Hard of hearing (D/HH) individuals face daily challenges, with over 48 Million living in the United States [4, 12]. For many D/HH individuals, sign language is their first and primary language. Despite the critical importance of sign language, the widespread shortage of interpreters often results in barriers in achieving functional equivalency in various aspect of daily life. These challenges are pronounced in social interactions [10], educational environments[22], and even healthcare settings [42].

A common misconception is that sign languages are merely visual representation of the spoken language of a country, leading to a significant underestimation of the access problems faced by D/HH individuals. In the United States, for example, many D/HH individuals' native language is American Sign Language (ASL), not English. Importantly, ASL possesses a distinct form and structure from English, and cannot be considered a merely visual depiction of English. ASL has its' own grammar, morphology, phonology, and syntax [35]. However, despite these linguistic differences, when ASL services are not available D/HH individuals are often expected to navigate a hearing dominant world with insufficient tools. The average D/HH high school graduate in the US only has a 4th grade reading level in English but the communication alternatives available - such as closed captioning, typing/writing, or a teletypewriter (TTY) - are all English-centric. This systematic disconnect leads to pervasive barriers evidence by a 16% employment gap (54% of D/HH individuals employed vs. 70% of hearing individuals), a 15% education gap (18% of D/HH individuals

with bachelor's degrees vs. 33% of hearing individuals), and a 22.1% labor force participation disparity (42.9% of D/HH individuals not participating vs. 20.8% of hearing individuals) [18]. These statistics highlight the enduring impact of a world designed without functional sign language access, where D/HH individuals are often left to think outside the box when it comes to their communication access.

To ensure these significant gaps are minimized, it is crucial for accessibility to be made available within a sign language modality. As the U.S. Disability Advisory Committee highlights, "without the ability to have other participants' audio converted to sign language and to have their own sign language converted to speech, a person who is Deaf or Hard of Hearing [...] may not be able to effectively participate in video conferences [or conversations]" [8]. This shows the importance of integrating sign language services directly into everyday products and life. Currently, these services are provided through sign language interpreters. However, the demand of interpreters far exceeds the supply, often leaving critical situations without access to sign language services. For instance, in the US there are millions ASL signers, there are only 10,000 certified sign language interpreters [37]. This often leads to cost prohibited pricing with private interpreting services priced at around $2.20 per minute (with often a minimum booking length of 2 hours), and the Federal Communication Commission (FCC) rates ranging from $6.27 to $ 7.77 per minute. [7] These cost and the lack of interpreters are often cited as barriers to reaching a world where ASL is in every day product and life, leaving many D/HH individuals with limited access. Furthermore, the inflexible pipeline for training interpreters - requiring at least four years - renders the D/HH community vulnerable to shocks, such as the Covid-19 epidemic, which exacerbate the already-pressing interpreter shortage [9]. To address these challenges, there is an urgent need for innovative tools that provide D/HH individuals greater autonomy and more options for functional accessibility.

With the recent Artificial Intelligence (AI) advancement, there is a growing interest in developing and evaluating automated interpretation solutions powered via AI to help alleviate some of these issues. While much work has been conducted within the Machine Learning literature striving towards the development of these technologies [6, 19, 27, 30, 32], no research has yet tested automated interpretation systems with D/HH users to determine whether these systems can actually facilitate a conversation in real-world scenarios. Similarly, within the Human-Computer Interaction (HCI) literature, only closed-domain with limited capabilities in translating a handful of signs, have been studied. Therefore, guidance remains scarce surrounding the usability of open-domain translation systems. Additionally, all studied systems are one-way (only either facilitating communication from the D/HH individual to the hearing individual, or from the hearing individual to the D/HH individual). Due to this scarcity of research, no well-understood standards exist surrounding sign language technology, including a minimally accepted standard to ensure effective communication. Furthermore, there exists no literature comparing human interpreters with AI systems, necessitating the establishment of baselines for accuracy, efficiency and robustness.

To this end, we present an ML-powered bi-directional communication system allowing D/HH and hearing interlocutors to communicate, both using their primary language. Additionally, we perform several robust studies surrounding usability, real-world accuracy, user preferences, and efficiency. We perform a comparative analysis of commonly used human interpreter solutions to establish acceptable baselines surrounding some of these metrics. We find that our system is non-inferior to interpreters when it comes to accuracy, efficiency or robustness. We ultimately demonstrate that our system is acceptable for real-world use is many scenarios. We close by envisioning a hybrid future where interpreters and AI can integrate sign language into everyday products and life. In total, we run four user studies (ranging across 106 D/HH individuals). By disseminating this research, we hope to establish evidence-based minimum acceptable standards, baselines, and targets for future work.

To the authors' best knowledge, this is the first study on the real-world utility of a bi-directional system. Therefore, this paper's key contributions are as follows:

- **We present** an ML-powered bi-directional system capable of automatically translating sign language using AI via SLR and SLP.
- **We run** robust user studies to demonstrate efficacy and establish a 'baseline' usable system.
- **We perform** the first comparative analysis between human interpreters and bi-directional communication systems to situate our technology within the context of interpreter accuracy and efficiency.

Additionally, this is the first study analyzing the actual utility and usability of such automated systems.

2 Related Work

This section examines prior research relevant to the central themes of the current study and outlines the motivations behind our key innovations. This study designed an AI-driven bi-directional system consisting of two major components: sign language recognition (SLR) and sign language production (SLP). Therefore, Section 2.1 discusses prior research into Sign Language Recognition (SLR). Section 2.2 introduces previous research on Sign Language Production (SLP) via signing avatars. Finally, Section 2.3 reviews the background of system usability in this domain, including aspect of resource efficiency that it addresses.

2.1 Sign Language Recognition (SLR)

SLR refers to the use of computer vision and natural language processing techniques to interpret and understand sign language automatically. It is a growing field within HCI aimed at narrowing the communication gap between the D/HH community and hearing individuals for various contexts, such as facilitating communication with care providers [16], learning mathematics [1], and enhancing accessibility in public services (e.g. banks) [27].

Recent advancements in machine learning (ML) and natural language processing (NLP) have fostered significant interest in developing automated SLR systems that can facilitate real-time translation between sign language into spoken or written language for the D/HH community. For example, Rastgoo et al. [30] proposed a real-time isolated hand sign language recognition model that included a single shot detector, 2-dimensional CNN, singular value decomposition, and Long-short Term Memory (LSTM) to extract and process discriminative features from 3-dimensional hand key-points. They confirmed that the model achieved competitive results in both accuracy and recognition time on four benchmark datasets (e.g., RKS-PERSIANSIGN (99.5 ± 0.04) [29]), demonstrating its efficiency. Lee et al. [19] designed an application that included a leap motion controller for real-time ASL recognition in a whack-a-mole game format to improve the effectiveness of ASL learning. A LSTM combined with k-Nearest-Neighbor was used to classify static and dynamic ASL signs based on extracted features such as finger angles, distances, and sphere radius. Results showed that the model achieved an average recognition accuracy of 91.82%. Sharma et al. [32] employed a 3D Convolutional Neural Network-based model to recognize dynamic hand gestures in ASL from volumetric video data. They demonstrated that the proposed approach outperformed the existing state-of-art models (i.e., 3.7% improvement in precision).

Notably, studies which examine human factors of these models only test constrained, closed-domain systems that cannot translate arbitrary signed strings. Alternatively, some studies evaluated isolated systems that can recognize only fingerspelled words or individual signs, rather than continuous strings of signs. Studies which report continuous, unconstrained SLR systems do not conduct human subject testing to validate their models work beyond a validation set testing. Specifically, these systems have not been tested with actual D/HH signers, evaluated in real-world scenarios, or

had their human factors analyzed. Very few attempts have been made to understand the feasibility of using such an unconstrained automated SLR system in the wild.

To this end, we developed and studied an open-domain continuous real-time SLR system as a practical tool for D/HH users in daily life. Given the shortage of prior research in this critical area of HCI, the study contained in this paper will explore a relative comparison of SLR systems and preexisting interpreting infrastructures. Not only will our study discuss the relative usability of our SLR system, we also aim to establish a baseline to compare its relative accuracy and efficiency when contrasted with human interpreters.

2.2 Sign Language Production (SLP)

SLP refers to the use of avatars or other mediums to generate digital representations of human figures to visually depict sign language [2]. These models are typically animated 3D models used to replicate hand movements [6], facial expressions [15, 41], and body language [17, 33] necessary to communicate in ASL.

Advancements in AI, particularly in consistency and flexibility, have significantly expanded the use of avatar rendering in ASL applications. These avatars, typically depicted as either cartoonish or human-like figures, have gained widespread acceptance and shown promising outcomes [39]. For instance, Quandt et al. [28] developed the Signing Avatars and Immersive Learning (SAIL) system, which utilizes the LEAP Motion system to track gestures within a virtual reality setting aimed at teaching ASL. Usability testing of SAIL revealed that users had a positive overall experience and expressed favorable feedback regarding the system's potential for ASL education. Similarly, Xu et al. [40] employed a transformer-based Conditional Variational Autoencoder to generate ASL fingerspelling alphabets. They tested this approach on three prevalent video-based human representations: two-stream inflated 3D ConvNet, 3D body joint landmarks, and body joint rotation matrices. Their findings indicated that the most effective generation of ASL alphabet signing was achieved using rotation matrices of the upper body joints and the signing hand. Additionally, Baltatzis et al. [3] introduced a diffusion-based Sign Language Processing (SLP) model that trains human-like characters on a large-scale dataset comprising 3D dynamic ASL sign sequences paired with text transcripts. They demonstrated that their method significantly outperformed existing SLP techniques in generating dynamic 3D avatar sequences from diverse discourse domains. This was accomplished by utilizing a diffusion process on an anatomically informed graph neural network based on the SMPL-X skeleton [25].

The utility of these SLP avatar rendering in a communication system remains untested. To this end, we present our generative SLP avatar and conduct usability tests to determine the practical implications of deploying such a system in a communication modality. We contend that a combined SLR and SLP system will be more usable for D/HH individuals, boosting the sustainability of their interactions and further enhance their functional equivalency with their hearing counterparts. The system will then be contextualized within resource management in our study.

2.3 System Usability

System Usability refers to the measurement of how easy and efficient a system is to use [38] by users. It is a critical factor in the adoption and effectiveness of SLR and SLP systems, especially when they are intended for use by the D/HH community to facilitate communication with hearing individuals. Several key factors contribute to the success of both system applications' usability:

- **Ease of Use** An intuitive system interface with a minimal learning curve is crucial for the effective adoption of SLR and SLP systems. It should allow simple navigation among system components, customize system

settings, and offer accessible feedback mechanisms for users [31]. Users who are unfamiliar with systems should be able to interact with them easily, especially for members of D/HH community, ensuring that the system accommodates their preferences for visual communication. However, most of existing communication assistive technologies to facilitate D/HH community in communication have still remained in the prototype phase [10] except HandTalk, which is publicly available software for translating text or audio to ASL or Brazilian Sign Language (LIBRAS) [14].

- **Real-time Performance** The ability to provide low-latency, real-time SLR and SLP is paramount for communication in the D/HH community, especially in live settings. It requires the bi-directional system to translate signs to text/speech or text/speech to signs with minimal delays and maximal accuracy that need to be carefully balanced, as a unreasonable lag in translation could disrupt the natural flow of conversation and reduce user satisfaction [23]. For example, Jolly et al. [36] collected narratives from 15 D/HH college students across the United States regarding their experiences with real-time captioning services. Their findings confirmed that real-time captions were an effective tool in helping D/HH students access to information and facilitate communication in classroom environments even they might struggle to overcome initial barriers (e.g., lag).
- **System Reliability** Reliable systems must interpret well both from ASL (SLR) and into ASL (SLP). Additionally, reliable systems should maintain high accuracy across diverse signing styles and contexts, ensuring consistent performance despite variations in signing speed, regional dialects, or individual user preferences. They should be robust to environmental factors such as background noise, lighting conditions, and occlusions that may affect gesture recognition and avatar rendering. Furthermore, reliable systems must handle incomplete or ambiguous inputs gracefully, providing meaningful interpretations and corrections when necessary.

Despite the critical nature of these factors, very few studies have explored the usability of SLR and SLP systems in real-world applications. The scarcity of comprehensive usability studies in day-to-day contexts has limited the understanding of how these systems would perform under real-world conditions. This lack of practical insights hinders the development of truly accessible and user-friendly SLR and SLP technologies. Hence, our study aims to address the gap using a thorough usability test of both systems in real-life settings, ensuring the design of SLR and SLP systems can facilitate daily communication effectively between the hearing and D/HH communities.

Usability is crucial when thinking of addressing resource constraints in the realm of SLR and SLP bi-directional systems. The response time for interpreter services can vary, typically requiring time to pick up when called or even planning in advance. This not only limits access for D/HH individuals but also places an undue burden on organizations seeking to provide timely communication. In contrast SLR and SLP bi-directional systems can help address not only the shortage of interpreters, but also the logistical time challenges associated with interpretation services should that the bi-directional system have both low latency and high accuracy.

Effective resource management, particularly when talking about time, yields positive externalities for both marginalized communities and their stakeholders as a whole. By enhancing usability in a bi-directional system, we can sustain conversations, increase efficiencies, and promote adoption within the D/HH community. A system that is user friendly will allow for seamless experiences that increases autonomy of D/HH individuals by giving them a tool they can use anywhere at anytime. However, if the bi-directional system lack usability, it cannot be deployed to mitigate the time delays associated with traditional communication methods, rendering it ineffective in real world scenarios.

3 Bi-directional communication system

AI-powered automated interpretation system should be bi-directional allowing the D/HH individual to communicate with a hearing individual and vice-versa. Such bi-directional systems would enable seamless two-way communication, ensuring that both D/HH and hearing individuals can fully engage, express, and understand each other without communication barriers. While one-way systems such as SLR or SLP addresses translations in either direction, they do not fully support natural and interactive communications.

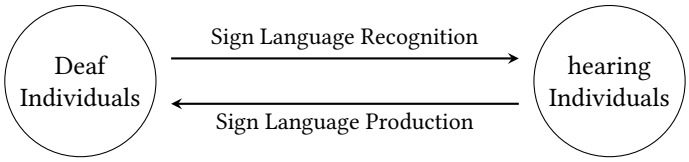

Fig. 1. An overview of system architecture

In addition to being bi-directional, we posit SLR and SLP systems have their own requirements. The SLR system should be open-domain, meaning that there are no fixed domain limits (i.e., the system should not be constrained to one topic area such as weather). Additionally, the SLR system should be able to understand a variety of ASL constructs (e.g., topicalization, indexing, lexicalized fingerspelling, and at least minimal understanding of ASL Classifiers). The SLR system should naturally also be continuous, and not restricted to isolated words, phrases, or fingerspelling. Finally, the SLR system should be able to capture the wide diversity of ASL.

The SLP system should be trained on actual signing data, and learn to translate (rather than just use hardcoded keypoints). The SLP system should not use hard-set rules (e.g., applying topicalization), as language is far more fluid.

3.1 User experience and interface

Our bi-directional interface (as shown in Fig. 2) contains several distinct components to ensure smooth communication.

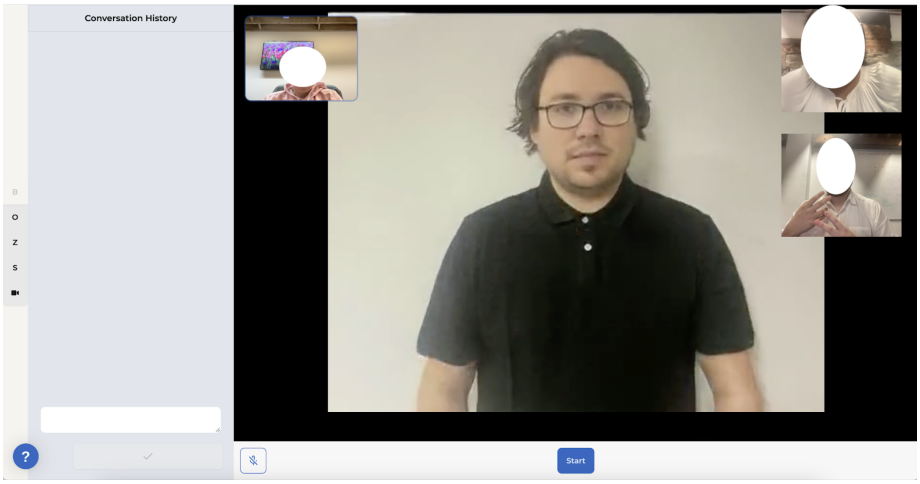

Fig. 2. An overview of bi-directional interface design. A brief demonstration can be found in the supplimental material.

3.1.1 Sign Language Recognition Interface. For the Sign Language Recognition, we have a user self view, so they can appropriately adjust their positioning, framing, and self adjust. Additionally, we have a button which changes state between “Start", “Stop", and “Redo" used to trigger the sign recognition engine to begin recording a sentence and stop recording a sentence (to send it for processing). Future work will examine methods of removing the start/stop button to allow for continuous streaming sign recognition. After the user concludes their signing, the button changes to a “loading" icon to inform the user that their signing is currently being processed. Approximately one to two seconds later, the system finishes and produces a prediction in the text box. Finally, the “start" button changes to a redo. Upon the text showing up in the text box, the user can either:

- **Submit** the text by pressing the check mark.
- **Edit** the text by typing into the text box.
- **Redo** their signing the pressing the "redo" button.

Sign Language Avatar Interface. When a hearing person speaks, the avatar is automatically prompted to sign whatever the hearing person said. While the avatar is translating, a “Translating..." icon shows in front of the avatar. Unlike the Deaf user, the hearing person has limited ability to start, stop, or repair their output. We decided to do this as in actual deployments, this will be deployed by having the tablet face the Deaf individual. Therefore, the hearing individual will not have the ability to see or interact with the system. We found this to be an acceptable trade off, as speech recognition performance is well understood and speech streaming recognition is well established. We can therefore presume both high accuracy and fidelity The hearing (or Deaf) individual does have the ability to mute the microphone by pressing the mic button.

History View. In addition to the avatar and sign language recognition, we additionally surface a history view containing the entire conversational history. This is done in effort to remediate cognitive fatigue that often occurs in Deaf/hearing communication.

3.2 Machine Learning Models

To do this, we begin by collecting a paired corpus of ASL. From this, we proceed to a three feature extraction steps:

- **Annotation of Gloss**: We annotate each sequence with the gloss it contains.
- **Annotation of Linguistic Information**: We annotate each gloss and each sequence with linguistic (primarily phonetic) information. We drive this annotation via an AI-assisted linguistic annotation tool. We dimensionality reduced the linguistic information to $\mathbb{R}^{K_1}$.
- **Extraction of low-dimension data representation** We extract a low-dimensional feature representation from each data point containing pose, ResNet features, and cropped areas of interest (face, hands).

From this, we obtain dataset $D = \{(l, x, g, e, f, l)_i\}_{i=1}^{N}$ for linguistic, low dimensional representation, pose, gloss, and English features respectively. Examining each space (let T_e be the number of tokens in the English translation T_a be the number of frames in ASL, and T_g be the number of tokens in gloss):

- $l \in \mathcal{L}^{T_a} \subset \mathbb{R}^{T_a \times K}$ represents strings of linguistic phonemes.
- $x \in X \subset \mathbb{R}^{T_a \times K_2}$ represents the low-dimensional data representation, containing a concatenation of extracted pose, extracted ResNet, and dimensionality-reduced per-frame cropped regions of interest. The per-frame

cropped region of interest is important as motion blur caused by the high-velocity motion in signing frequently prevents the pose prediction model from functioning[1].

- $g \in \mathcal{V}_G^{T_g}$ represents the strings of gloss. Note that we primarily rely on L to capture the broad morphological[2] variations of root-signs.
- $k \in \mathcal{G} \subset \mathbb{R}^L$ extracted signing style attributes such as signing speed, signing space, and lexical choice.
- $e \in \mathcal{V}_E^*$ represents strings in the vocabulary of English. Note that these are not words, but rather tokens.
- $f \in F* \subset \mathbb{R}^{K \times K^*}$ represents videos composed of K by K frames

3.2.1 Sign language Recognition. For Sign Language Recognition, we aim to map from ASL videos to English. As we have limited data and raw videos are of exceedingly high dimesionality ($\mathbb{R}^{T_a \times 512 \times 512}$ for 512 x 512 videos), we perform dimensionality reduction. First, we limit ourselves to extracted features $x \in \mathbb{R}^{T_a \times K_2}$ where $K_2 << 512 * 512$. We perform a fusion over several feature sets to capture salient information from signing videos. In particular, we run a pose recognition model across the input video to obtain keypoint coordinates. The pose recognition features alone are insufficient, however, as motion blur induced by fast signing renders pose recognition models non-functional.

We therefore include ResNet features, and learned features from croppings of both hands and the face as inputs into the fusion model. By equipping the feature extraction fusion model with both coordinate information and cropped images of hands, our model should in theory be able to recover trajectory information and perform the associated learned deconvolution to deblur the image prior to performing its own keypoint extraction. If we equip the feature extractor with only pose information, keypoints lost due to blurred images would be unrecoverable due to information lost during the keypoint identification process process. Our feature extractor therefore accepts the pose information in embedded in $\mathbb{R}^{T_a \times 2P}$ where P is the number of keypoints extracted by the pose predictor, ResNet-50 final features embedded within $\mathbb{R}^{2048}$, and three images of cropped regions surrounding the hands and face (10% margin outside of any detected keypoint, or a fixed margin surrounding the globally positioned hand if no hand keypoints are detected) situated in $\mathbb{R}^{128 \times 128}$.

Following this, we perform a sign boundary detection. We begin by extracting the 'sign index' for each sign annotated within a given sentence. Therefore, this target would be $b_i \in \{0, 1, ..., W\}$ (assuming at most W words could occur per sentence). Following this, we use $h_\phi : X_{T_a} \rightarrow \mathbb{R}^{W \times T_a}$ which accepts our low dimensionality feature space and predicts the sign index (we assume that there are always less than W signs in a given region). h_ϕ is modeled as a transformer model. $\mathcal{L}_{sb} = CCE(b, h_\phi(x))$.

Next, we perform individual sign predictions W times. Our sign predictor z_ψ regress to our gloss and linguistic space $\mathcal{V}_G^{T_g} \times \mathcal{L}^{T_a}$. Specifically, for each sign $0 \leq i < W$, we compute an attention mask by examining the sign boundary prediction $\hat{b}$. This model is designed so that z_ψ evaluation on the i'th attention map produces predictions for the i'th sign. We use categorical cross entropy for the glosses and MSE for the lingusitic features: $\mathcal{L}_l = \sum_i \mathbb{E}[(l_\psi(b_i^T x)_0)^2] + CCE(g_i, l_\psi(b_i^T x)_1)$. We perform smoothing via a language model.

Finally, we desire to translate the gloss and linguistic information. To do this, we simply use a transformer t_ζ trained on the gloss. This translation model accepts the linguistic and gloss predictions and predicts the English tokens. This is trained using categorical cross entropy $\mathcal{L}_t = CCE(t_\zeta(\hat{g}, \hat{l}), e)$ where $\hat{g}, \hat{l}$ are our predicted gloss and logits.

[1] linguistically, as handshapes are frequently changed during Movement portions (see Movement-Hold model), it is crucial to capture the handshape changes

[2] while this set only captures the phonologic details, as morphemes are representable as sets of sequences of phonemes, it serves dual role to carry morphological data through the system.

As these models were cascaded, this provides ample inductive bias to prevent overfitting due to domain complexity and limited data. Using this model, we achieved ScareBLEU of 55.1 on our test set.

3.2.2 Sign language Production. For our sign language avatar, we utilize a multi-step approach to generate our human-like avatar.

Generation of pose information from text. We begin by regressing from English to a written sign language analog. This step was extracted from the pose generation as the output contains the most variance (i.e., by fixing the sign language produced, the actual production of motion sequences becomes substantially less variable).

We learn $f_\phi : E \times \mathcal{G} \rightarrow G \times L$ via f_ϕ (the gloss and linguistic features). Note that we not only regress to the gloss, as gloss alone typically omits many crucial morphological features. Additionally, note that we provide global linguistic features to allow us to provide guidance to our models on signing-style. This model can be done via a Seq2Seq model such as a transformer, and trained via cross entropy (as both the gloss space and linguistic phonetic space are discrete). Our loss function is as follows: $\mathcal{L}_l = (f_\phi(e,k)_1 - l)^2 + CCE(g, f_\phi(e,k)_0)$.

Generation of pose information from written sign analog. Following this, map from our gloss-linguistic written sign analog space to actual pose data. We aim to map $g_\psi : G \times L \rightarrow X$. This can be done easily by parameterizing a Transformer with ψ. To obtain continuous outputs, we regress to two distinct values: the actual pose p_i and a terminator signal T. If $\hat{T} > 0.5$, then we terminate the generation of poses. We use a BCE loss function over T paired with an MSE over X. $\mathcal{L} = BCE(g_\psi(g,l)_0, T) + (g_\psi(g,l)_1 - X)^2$

Generation of frames from pose information. Our sign language avatar uses a deepfake to generate a video of a desired individual signing. In particular, it uses a conditional general adversarial network conditioned on pose information (similar to the Everybody Dance Now work)[5]. In particular, we unroll frames and extracted features (f, x) into frames and their corresponding features. We then train a generative adversarial network (GAN) using a generator G_ρ and D_θ. Through this, we thereby learn the inverse problem to low-level feature extraction. We choose GANs (over diffusion based models) in particular for speed. While diffusion based models have been able achieve good image fidelity, due to their inference requiring the repeated evaluation of a function due to their Markovian nature, they tend to be at least one order of magnitude slower than a GAN.

4 Testing Methodology

As our primary goal is establish baselines and contextualize our work in the context of human interpreters. We have identified two broad research topics: the usability of the bi-directional communication system and a comparative evaluation of its accuracy. Each of these topics has its sub-questions.

- **RQ1:** Is the bi-directional communication system usable and well-received, and how do individuals perceive the accuracy and experience of interacting with the signing avatar and automated interpretation solution?
- **RQ2:** How do automated interpreting solutions compare to human interpreters in terms of efficiency, robustness, and accuracy, and would a hybrid system of AI and human interpreters improve overall accuracy?

4.1 Design of Experiments

We wish to observe and analyze the impact these models have on users, and perform comparative analysis against interpreters. Critically, if improperly designed, even accurate models can have counterproductive results when actually

deployed. As our system is comprised of three primary components (1-the user interface, 2-the SLR, 3-the SLP), we must measure each component individually, as well as when combined. To this end, we conduct four Institutional Review Board (IRB)-exempted usability studies to examine the efficacy of these models contextualize within our described user interface:

(1) **SLR: Accuracy Study (RQ2)** A comparative analysis was conducted with $N = 25$ individuals to assess the accuracy of our system in relation to that of Video Relay Service (VRS) interpreters.
(2) **SLP: Avatar Accuracy Study (RQ1)** A survey was conducted with $N = 50$ D/HH individuals to measure their sentiment and attitudes towards cartoon vs human-like signing avatar, including metrics such as accuracy and preference.
(3) **Combined System: Usability Study (RQ1)** A user study was conducted with $N = 18$ D/HH individuals to evaluate the usability and broad sentiment towards the bi-directional system (SLR and SLP).
(4) **Combined System: Efficiency Study (RQ2)** An efficiency study was conducted on $N = 13$ Video Relay Interpreting (VRI) calls to measure interpreter speed with conversation speed within our system.

The inclusion criteria for all studies were as follows: D/HH participants should be at least 18 years old, use ASL as primary language with sufficient proficiency, and should not have any conditions which would impact their ability to interact with our system (e.g. Cerebral Palsy, Deaf Blind). Interested D/HH individuals contacted the research for voluntary participation. Participants completed a screening process to determine eligibility. Eligible participants would be provided informed consent. The Deaf researcher thoroughly explained the study, potential risks, and benefits. Each participant also provided demographic information (e.g., age) and participants received $25 as compensation for their role in the study.

4.2 Analysis of SLR accuracy

We begin by analyzing SLR accuracy. Our goal is to evaluate SLR accuracy against established baselines. While, current literature typically performs its evaluation solely against their own test datasets, we aim to assess the real-world feasibility of deploying SLR. A meaningful metric, therefore, is to determine whether our SLR system is inferior to Video Relay Service (VRS) interpreters. We choose to use VRS interpreters as much interpretation received by a Deaf individual on a daily basis will be from VRS interpreters (rather than from the private sector). While an inferior system may still have some utility, a non-inferior system would clearly offer acceptable levels of accuracy for practical use.

Automatic ranking of machine and human translations can be challenging [20]. We therefore opt for a holistic approach, utilizing multiple metrics to score each sentences. In particular, we choose to use:

(1) **SacreBLEU** SacreBLEU is a standardized form of the prolific BLEU metric commonly used to measure translation accuracy. BLEU and SacreBLEU have been widely used since its inception as the de-facto measure of a translation systems' quality.
(2) **ChRF** ChRF is a relatively recent metric which computes character n-gram precision and recall (optionally with added word-level overlap). ChRF is often used as an auxiliary metric alongside word-level metrics such as BLEU and TER as it uniquely provides 'partial credit' when encountering morphological errors.
(3) **TER** Translation Error rate computes the percentage of additions, edits, or deletions that a source text must undergo to become the reference text. For convenience, we use $100\% - TER$ (so that higher is better). This corresponds to a 'Translation Accuracy Rate'.

We delineate accuracy as both the mean performance in aforementioned scores and the variance in performance. While an automated system may be marginally better or non-inferior with respect to mean performance, if that system experiences a significantly higher variance, it may still not be acceptable for deployment. We therefore wish to measure if (1) our system consistently under-performs against interpreters (2) or if our system has a higher variance. To measure inferiority in this context, we prepared a list of 500 random sentences (with no five-gram overlap with our training data). We partition the data into 25 equal partitions and ask Deaf individuals to sign these sentences into our system and to interpreters. Recruited Deaf individuals had a mean age of 36.7 years with a standard deviation of 11.6 years. Importantly, the same individuals signed the same sentences to both systems.

To streamline collection, we utilize VRS as a baseline. VRS is an FCC funded service provided to D/HH individuals to allow them to make phone calls through a communication assistant (CA) such as an interpreter. Several companies provide VRS services such as Sorenson, Convo, and Purple. These services make calls on the Deaf individual's behalf and interpret the call between the Deaf and hearing caller. Importantly, there are many restrictions to the usage of VRS, as their usage is restricted to telecommunication systems. Additionally, usage of VRS is prohibited if both callers are in the same room. These restrictions limit the efficacy of VRS to just a phone provider. We chose to use VRS over other available interpreting options as VRS systems are required meet certain minimum standards. Additionally, as VRS systems are used often, VRS represents a significant portion of the average Deaf individual's everyday interpretation. To facilitate this, the participants were provided with a script and instructed to call a phone number and leave a voicemail. Given that the script contained a series of unconnected sentences, interpreters were made aware up front that they should not attempt to connect the sentences together to form a story.

We chose to evaluate the system on isolated sentences as testing on longer pre-scripted interactions would be likely to introduce bias and confound the analysis of accuracy of the recognition system (as it would be unclear if an accuracy of a prediction was caused by contextual clues or the actual recognition engine). While interpreters perform tasks more complex than isolated phrase translation, it is unlikely that a interpreter or an interpretation system would be able to perform accurate in a high-context situation if they were unable to translate individual sentences. For an analysis of performance on longer in-context interactions, observe the combined system usability test presented below.

To capture and measure human interpreter accuracy, we setup a Twilio voicemail configured to record audio. Deaf participants were asked to call that voicemail and read off the script. The recording was then retrieved and automatically transcribed via speech-to-text. The transcript was then checked by a human annotator to ensure that the transcript accurately captured the precise syntax used in the recording. Finally, the resultant sentences were aligned with the prompted sentences. To capture and measure the automated system, we instructed Deaf individuals to sign the sentences into a video. That video was translated via our system, and had each sentence aligned with the original sentences. In both cases, the predicted translation and actual translation were stored with the Deaf participant's participant ID and the ground truth (of what was prompted).

As we are additionally interested in the feasibility of 'teaming' our AI model with an interpreter, we are interested to see if an ensemble model provided with both the AI output and an interpreter output would outperform either an interpreter or AI system. To this end, we ensembled the human interpreter predictions with the AI prediction via a process similar to LLM-Blender.

The prediction from SLR system, the human interpreter, and the ensembled prediction were all associated with the underlying ground-truth English used to prompt the Deaf individual.

4.2.1 Statistical Analysis Methodology. We wish to measure both the pairwise difference in performance between each system, and the variance of accuracy of all three systems (AI only, Human only, AI + Human ensemble). Note that for the comparison between an AI and Human interpreter, we wish to merely show non-inferiority, while with AI + Human ensemble vs Human, we wish to show superiority. To quantify non-inferior, we use the maximum un-noticeable metric difference described in [20] (we use the maximum value for which the probability of preference given a difference of at most that value does not exceed .5). The margin of non-inferiority can be found in the table below.

	CHRF	TER	BLEU
Maximum Unnoticeable Difference	1.0	1.5	3.0

Table 1. Margin of Inferiority based on Maximum Unnoticeable Differences

Unfortunately, for both of these test sets, our actual performance is not i.i.d. In particular, each of the 25 sentences signed by a single individual is likely to have correlation with each other. Failing to account for this dependence greatly inflates the significance. To cope with this, we utilize Linear Mixed Effect model with a random effect per-participant. To perform mean accuracy analysis, we use Generalized Estimating Equations (GEE) to estimate the pairwised mean difference of scores when the same sentence is predicted by two different systems (e.g. compare human interpreters against the ensembled prediction). For GEE, we use our participants as our clusters, and use an exchangeability correlation structure. We perform this analysis for each accuracy metric. To actually run our hypothesis tests, for non-inferiority, we use thresholds for each score which are unnoticeable[3] as the margin for inferiority and run standard non-inferiority tests.

Variance of Accuracy Analysis. Although our difference in score distributions are normal, the actual score distributions are highly non-normal. Running any kind of null-hypothesis rejection test would require us to contend with both the non-parametric nature of the the difference score distribution and the hierarchical dependence of our data. We instead propose running a non-parametric bootstrap and utilizing the resultant distributions to construct confidence interval intervals for variances of interest. Those intervals can be used to perform confidence-interval based hypothesis testing. First, we want to examine the score at cluster level (i.e., the variance of the system is more variable per-user. We additionally want to examine intra-cluster variance (perhaps overall the variance between users is non-inferior, but within an individual, variance of accuracy is higher).

Therefore, for this study we wish to answer:

- **Accuracy** Does any accuracy metric find our accuracy inferior to interpreters.
- **Robustness** Is the variance of any of our accuracy metrics inferior to interpreters.

D/HH members distribute themselves across a variety of signing styles. In particular (note that in an abuse of terminology, we refer both both of these as ASL), some individuals use a fairly linguistically pure ASL mostly unaffected by English, while others use the same signs but is similar to English in word order and can even appear nearly identical, such as in Signed Exact English (SEE). Rarely do individuals actually occur at either extreme. It is typical for individuals to fall somewhere along a spectrum between pure ASL and SEE. As this is a pidginified form of ASL, this is commonly referred to as Pidgin Signed English [4]. This gradient has historically been used to perpetuate privilege and bias within

[3]We define the unnoticeable threshold as the maximum difference in scores between systems where no more than 50% of evaluators would perceive one system as superior to the other.
[4]To be precise, PSE is a linguistically incorrect term as PSE has since formed it's own acceptable grammar and thereby been creolized and is not a Pidgin.

the D/HH community through linguistic prescriptivism [34]. We therefore take a linguistically neutral and descriptive approach to avoid further marginalization.

Individuals are likely to disagree with other individuals across the ASL-SEE spectrum in terms of 'correct signing' (as one person may consider SEE to be correct, while another may consider ASL to be correct). To avoid perpetuating bias, and to examine this effect, we opt to offer two avatars conditioned on signing style (an avatar leaning more SEE and an avatar leaning more ASL).

We hypothesize that while preference may exist between the avatars, individuals who do not like one avatar are more likely to like the other avatar.

4.3 Avatar Accuracy Study

We wish to analyze user preference on SLP Avatars. Notably, signing D/HH members distribute themselves across a distribution of signing styles. In particular (note that in an abuse of terminology, we refer both both of these as ASL), some individuals use a fairly linguistically pure ASL mostly unaffected by English, while others use the same signs but in English word order with English word ending signing Signed Exact English (SEE). Rarely do individuals actually occur at either extreme; typically individuals are distributed somewhere between pure ASL and SEE. As this is a pidginified form of ASL, this is commonly referred to as Pidgin Signed English [5]. This gradient has historically been used to perpetuate privilege and bias within the D/HH community through linguistic perscriptivism. We therefore take a linguistically neutral and descriptive approach to avoid further marginalization.

Individuals are likely to disagree with other individuals across the ASL-SEE spectrum in terms of 'correct signing' (as one person may consider SEE to be correct, while another may consider ASL to be correct). To avoid perpetuating bias, and to examine this effect, we opt to offer two avatars conditioned on signing style (an avatar leaning more SEE and an avatar leaning more ASL).

We hypothesize that while preference may exist between the avatars, individuals who do not like one avatar are more likely to like the other avatar.

For our two avatars, we wish to evaluate:

- **Overall preference** Do users like the avatar. If so, is there a consistent preference?
- **Accuracy** Do users think that either signing avatar is accurate. If so, is there a systematic preference? How does this accuracy compare with reported accuracy metrics for interpreters?
- **Understandability** Do users think that either signing avatar is understandable. If so, is there a consistent preference?
- **Discordance** How much do users disagree with each other? Do users who dislike one avatar tend to like the other avatar?

To evaluate these questions, we collected responses from $N = 50$ D/HH individuals (Age (years): 30.5 ± 12.6) on our avatars using an online survey. The survey presents each avatar along with Likert-style questions. In particular, we show two samples per signing avatar: a long sequence (at least 3-5 sentences), and a short sequence (at most 1 sentence). The survey order is randomized to prevent order bias. Additionally, the sentences themselves were picked from a randomly picked pool of 10 long sentences and 10 short sentences. Both avatars signed each sentence picked. After each video, questions surrounding accuracy and preference were asked. Only after all avatar videos were watched and evaluated were the users asked to rate their avatar preference.

[5]To be precise, PSE is a linguistically incorrect term as PSE has since formed it's own acceptable grammar and thereby been creolized and is not a Pidgin.

To analyze overall preference, we can assume i.i.d. as all users are independent. To evaluate if a user likes an avatar, we threshold their Likert score for a 'liking' the avatar at a value of at least 'Slightly Agree'. We opted to perform an analysis using thresholded values (rather than mean and std scores) as users tended to provide extreme values which interfere with a normality assumption. We use these values to run a binomial test if there is a consistent preference for the avatar. For the per-avatar questions of accuracy and understandably, we average the response between both long and short videos so that we can produce a single overall score. These averaged scores are also i.i.d., so we can use an analysis similar to the overall preference. For discordance, we wish to demonstrate that acceptance rate of least one system provides a significantly higher acceptance rate than the individual acceptance rate of either avatar individually; if this occurs, we can suggest that there are a plurality of individuals who prefer exactly one system.

4.4 Combined System: Usability

While the previous two studies are focused on SLR and SLP, respectively, we are additionally interested in measuring overall usability of the combined system. Regardless of each sub-system, if the holistic system itself is not useful, individuals will not use the system.

To measure the combined system's usability, we run a study examining how effectively users can communicate with a hearing individuals with $N = 18$ Deaf and Hard of Hearing individuals Eighteen (Age (years): 30.85 ± 12.54). In particular, we propose using the system to facilitate remote conversations between D/HH and hearing parties. After users filled out the associated informed consent form and joined the call, the Deaf researchers played a video demonstrating the usage of the system. Following this, the Deaf researcher asked the user to choose a topic and have a brief conversation for 5-10 minutes with the hearing individual. Additionally, users were tasked with communicating with the same hearing individual by texting back and forth as a control. The treatment/control order was randomized per participant. Upon the conclusion of this conversation, participants were asked a series of subjective preference systems in which they needed to rank which system they felt they agreed that statement held more true for. Subjects were also asked a series of Likert based questions on the communication system. The experiment concluded with qualitative questions (e.g. please describe why you preferred system X; where would you want to use system X).

For this study we wished to primarily answer:

- **Preference** Do users like the communication system over their currently used communication mechanism (writing back and forth)?
- **Adoption** Would users be likely to use this system over writing back and forth if it were offered?
- **Satisfaction** Do users feel like this communication system meets their communication needs more then their current communication mechanism (writing back and forth)?
- **Efficiency** Did users feel like the system was more efficient than writing back and forth?

We analyzed all results using a ranking regime, so therefore could treat each outcome as a binary sample from a binomial trial. We therefore use a binomial test to test for significance for each question.

4.5 Combined System: Efficiency

We were finally interested in answering if our communication system was less efficient than the 'golden standard' of interpretation. In person, when interpretation is needed, commonly video relay interpreters (VRI) are used. This service calls an interpreter into an interaction, frequently via a tablet such as an iPad. VRI providers bifurcate their offering into pre-scheduled VRI (in which businesses need to schedule calls prior to the call occurring), and on-demand VRI (in

which businesses queue up and experience 'wait-time' prior to the interpreter joining the call). As most interactions are not prescheduled, we study the on-demand VRI interaction. We conduct a series of VRI calls to determine if carrying out a short interaction using VRI would be more efficient than using our sign language solution.

In particular, we select half of the sessions from the 'Combined system: Usability' and 'replay' conversations which occurred during the prior usability study. We measure the time from dialing the interpreter to the conclusion of the conversation. We want to determine what difference exists between the efficiency of VRI interpreters and efficiency of our system. We wish to determine:

- **Speed** Is our system slower by a noticeable margin than VRI interpretation?
- **Variance** Is our system speed less variable than VRI interpretation speed. If so, what explains the difference in speed?

To analyze speed, we merely subtract the time a conversation required in our system vs interpreters. As all differences were distributed normally, we use a paired t-test for non-inferiority to analyze speed. We deem a difference in conversation length of less than a minute as insignificant and therefore set our margin of inferiority to 60 seconds. For variance, the interpretation speeds were not normally distributed themselves. Therefore, we opt to perform confidence interval estimates via bootstrap.

5 Results

We proceed through each individual sub-study, followed by an overall analysis of answers to each of the research questions.

5.1 Analysis of SLR Accuracy

5.1.1 Evaluation and Comparison of Systems. In this section, we compare the performance of the AI, Human (VRS), and Hybrid systems using three evaluation metrics: BLEU, TER, and CHRF. Table 2 presents the mean (MU) and standard deviation (STD) for each system on these metrics.

Table 2. Mean and Standard Deviation of Evaluation Metrics per System. Higher is better.

Metric	System	Mean	Standard Deviation
BLEU	Hybrid	**59.67**	33.55
	AI	50.71	33.37
	Human	27.56	26.32
TER	Hybrid	**76.47**	40.25
	AI	64.61	31.94
	Human	30.88	45.50
CHRF	Hybrid	**76.74**	26.91
	AI	66.49	26.52
	Human	48.19	26.85

5.1.2 Statistical Comparison of Systems. To assess whether the AI and Hybrid systems perform at least as well as human interpreters, we conducted statistical tests. The p-values demonstrate that the AI system is not worse than human interpreters across all metrics, and the Hybrid system significantly outperforms both AI and human interpreters.

Table 3. P-values for Comparisons Between Systems. All results are significant.

Metric	Comparison	P-value
BLEU	AI not worse than Human	<0.001
	Hybrid better than Human	<0.001
	Hybrid better than AI	0.048
1-TER	AI not worse than Human	<0.001
	Hybrid better than Human	<0.001
	Hybrid better than AI	0.018
CHRF	AI not worse than Human	<0.001
	Hybrid better than Human	<0.001
	Hybrid better than AI	0.014

5.1.3 Analysis of Variance Between and Within Clusters. We further analyzed the variance between and within clusters using bootstrapped standard deviations. Tables 4 and 5 provide the inter-cluster and intra-cluster standard deviations, respectively, along with their associated standard errors. These values help illustrate the consistency of the system performance both across and within individual users. Recall that clusters represent individuals; intra-cluster variance represents variability of scores within that individual's samples, while inter-cluster variance represent variance between individuals. For convenience, we report inter-cluster and intra-cluster standard deviation.

Table 4. Inter-Cluster Standard Deviations measuring the deviation from user to user

Metric	System	Inter-Cluster STD ± SE
BLEU	Human	18.51 ± 2.74
	AI	12.27 ± 1.54
	Hybrid	10.91 ± 1.20
TER	Human	22.18 ± 2.91
	AI	13.85 ± 1.93
	Hybrid	15.27 ± 1.94
CHRF	Human	18.45 ± 2.77
	AI	12.22 ± 1.57
	Hybrid	10.89 ± 1.22

Table 5. Intra-Cluster Standard Deviations measuring the deviation within samples from one user

Metric	System	Intra-Cluster STD ± SE
BLEU	Human	19.07 ± 0.55
	AI	21.47 ± 0.55
	Hybrid	22.57 ± 0.55
TER	Human	35.90 ± 1.72
	AI	25.69 ± 0.85
	Hybrid	32.56 ± 1.50
CHRF	Human	19.08 ± 0.56
	AI	21.47 ± 0.54
	Hybrid	22.57 ± 0.55

Recall we estimate intercluster and intracluster standard deviation using a confidence interval paradigm. From this, we conduct appropriate confidence interval based hypothesis tests. Intercluster variance for AI is non-inferior to that of Humans (as we require the observed effect size is noticeable as described in Table 1) at $p < .05$, while intracluster variance is inconclusive. Note, however, that for intracluster effect size for human to AI, the effect size is close to or below the unnoticeable difference threshold presented in Table 1. Additionally, the effect of this small difference is minimal on user experience due to the disparity in mean metric values. These results suggest that our model is robust and is no worse than humans even on challenging samples.

5.1.4 Conclusion. From the metrics presented above, we conclude that our AI-based system is **not inferior** to human interpreters in terms of accuracy and variance, as demonstrated by the p-values for non-inferiority testing. Additionally, our hybrid AI+human system significantly outperforms both human interpreters and the AI system alone. This suggests that our hybrid system offers a promising solution that combines the strengths of both AI and human interpreters for superior performance in sign language recognition and interpretation tasks.

5.2 Avatar Accuracy

Table 6. Percent of users who agree with a given quality of SEE Avatar, ASL Avatar, and Either Avatar is true across various metrics.

Metric	SEE Avatar	ASL Avatar	Either Avatar
Accuracy	64%	76%	78%
Likeability	62%	76%	88%

Table 7. Percent of users rating a given system the highest under various metrics

Metric	Prefers ASL	Neutral	Prefers SEE
Accuracy	46%	26%	28%
Likeability	46%	30%	24%

The distribution of observed survey responses can be seen in tables 6 and 7 with table 8 containing p-values for significance (when tested with an alternative hypothesis that the mean of the given metric is at least 4 out of 6). We

Table 8. P-values that users rate a given metric higher for the ASL avatar than the SEE avatar.

Metric	P-value
Accuracy	0.0042
Likeability	0.0006

observe that overall users are positive towards both avatars, with the ASL avatar being preferred to the SEE avatar. We furthermore observe that users tend to prefer at least one avatar more than they prefer the other avatar (formally, if the a_i is the score for the i'th individual on the ASL avatar and s_i is the score for the s'th individual on the SEE avatar, $\mathbb{E}[\max(s_i, a_i)] > \max(\mathbb{E}[s_i], \mathbb{E}[a_i])$). The p-value for the superiority of the merged choices over each individual choice can be seen in Table 8.

From this, we conclude that individuals overall like both avatars (76% for the ASL avatar and 62% for the SEE avatar), and have a preference towards the ASL avatar. In addition, we conclude that individuals view both avatars as accurate (76% for the ASL avatar and 64% for the SEE avatar) with 78% rating that at least one avatar was accurate. These metrics align well with reported accuracy metrics for ASL interpreters of 72.7% [21]. Finally, we observe that a significant number of individuals (38%) like exactly one avatar. And for each avatar, there are individuals who prefer the other (as the likeability score for either avatar is higher than the likeability score for the ASL or SEE avatar with $p = 0.0027$), reinforcing the need for both avatars.

From the metrics presented above and the reported accuracy of human interpreters we conclude that our AI avatar is **non-inferior** to human interpreters, and provide a **viable alternative** when an interpreter is not otherwise available. In addition, we observe a **generally positive sentiment** towards these avatars.

5.3 Analysis of Usability

Table 9. Percent of users who agreed with each quality metric (Agreed) for the bidirectional system and percent who preferred (Preferred) the bi-directional communication system over writing back and forth.

Question	**Preferred**	**Agreed**
This system was easy to use.	88.89%	94.44%
This system met your communication needs.	86.11%	94.44%
If this system was offered to me (for example, with friends or family, or in a hotel check-in), I would use it.	85.19%	88.89%
This system was efficient.	79.17%	94.44%
Overall, I like this communication system.	78.89%	88.89%

Table 9 shows the percent of people that rated the bi-directional communication system higher for each question (with the p-value representing the alternative hypothesis that at least 50% of individuals preferred the bi-directional communication system). We observe that individuals broadly consider the bi-directional system superior to writing back and forth (88.89%), and would use that system if it were offered (85.19%); additionally, we observe that individuals felt that the system met their communication needs (86.11%) and was more efficient than writing back and forth (79.17%). Qualitative interviews following the experiment found all users who preferred typing back and forth stated their

preference was due to ingrained usage patterns. Additionally, half of the users who preferred writing back and forth communicated that they would have preferred our system had the conversation been longer or more in-depth.

Overall, users communicated that they preferred the 2-way system as they could communicate in their first language (they mentioned that using English requires them to translate in their head and is typically less natural) and voiced their interest in having this system unilaterally deployed everywhere. Additionally, users communicated this strong preference was in part due to the frustrating and slow experience of texting back and forth. Many users concluded with optimism, as this technology was described as akin to voice recognition.

5.4 Analysis of Efficiency

The average and standard deviation of times to hold a conversation are reported in Table 2. We observe that our system is no slower than VRI interpretation (and observe with significance via the paired t-test that our system is not noticeably slower with $p = 0.038$), with VRI interpretation having a significantly higher variance. We additionally observe that VRI interpreters have a significantly higher variance in the time to pick up the call. Wait time accounts for a majority of the variance within the VRI call length ($R^2 = 0.63$); interestingly, however, even when discounting wait time, the residual 37% of variance ($\sigma_r^2 = 2.87$ with σ_r^2 representing residual variance) still exceeds our system's variance in conversation time.

Table 10. Mean, standard deviation, and standard error of call length for AI and human. The unit is minute.

Subject	Mean	SD	Standard Error of SD
AI	5.05	1.02	0.16
Human	5.54	2.79	0.46

A bootstrapped analysis of the VRI call length standard deviation and AI interaction length standard deviation is presented in Table 10 (showing both the estimated value and the standard error). As the 95% confidence intervals for AI is [3.01, 7.08] and for VRI is [-0.03, 11.12], we conclude with $p = .05$ significance that AI call is less variable than VRI call length.

5.5 Research Questions and Standards

Revisiting our research questions:

RQ1 - utility and usability. We found that our system was indeed usable with overwhelming positive sentiment. Individuals liked our signing avatars. We found that to have broad appeal, it is advisable to use at least two avatars: one for ASL and one for SEE. Qualitatively, users have overall positive experiences.

RQ2 - comparison. We found that overall our system was no worse than interpreters in time, robustness, or efficiency. We additionally, found that a hybrid system ensemble both interpreter and AI predictions performs better than either individually. Based on these results, we find that:

- Bi-directional communication systems show much promise in real-word environments.
- To guarantee usability of an SLR system, the real-world BLEU score should be at least 50[6].
- To ensure acceptance of a signing avatar, multiple variants should be prepared for various signing styles.

[6]the minimum acceptable BLEU score may well be lower, but we only evaluated a 50-BLEU system

We additionally advise that work is focused in bi-directional systems rather than one-way systems. Finally, we reiterate that researchers involve key stakeholders from the D/HH community early, both in design, implementation, and evaluation to ensure acceptance; merely building for the D/HH community is insufficient.

The results from our four user studies underscore the promise of AI-powered sign language technologies in addressing key challenges faced by the Deaf and Hard of hearing (D/HH) community. Specifically, we found that the bi-directional communication system we developed demonstrated non-inferiority to human interpreters in terms of both accuracy and efficiency, suggesting that this technology can effectively bridge communication gaps. Our usability studies revealed overwhelmingly positive sentiment toward the system, with users preferring to communicate in their primary language, ASL, over typing in English. The introduction of multiple signing avatars also highlighted the importance of catering to diverse signing styles, ensuring inclusivity across a range of users.

These findings hold significant implications beyond mere technical performance. The ability to offer an accurate and user-friendly system for real-time communication has the potential to alleviate the long-standing barriers that D/HH individuals face daily. By examining these results in the context of broader social and psychological challenges, the following discussion explores how such technology can transform communication autonomy, independence, and emotional well-being for the D/HH community.

6 Conclusions and Discussions

Research into Deaf accessibility is crucial due to a lack of adequate accessibility tools currently available. For Deaf and hard of hearing (D/HH) individuals, navigating the hearing world can often be an isolating and frustrating experience. This is primarily due to limited opportunities for independent autonomous communication. While it is commonly held that access to technology and resources such as texting, emails, or even lip-reading fully bridge the gap, reality is far more complicated than those notions would suggest. Communication in a preferred language—such as American Sign Language (ASL)—is a vital part of expressing thoughts, ideas, and emotions for D/HH individuals [24]. Yet, many barriers to using ASL in day-to-day interactions remain significant.

Without access to sign language interpreters, visual aids, or ASL-fluent service providers, D/HH individuals are often forced to rely on written notes or gestures—methods that strip away nuance and lead to misunderstandings, particularly due to the differences between ASL and English grammar. Such limited communication options not only diminish independence but also foster a reliance and dependency on family members, friends, or interpreters to help navigate routine tasks [13]. Such a situation undermines an individual's autonomy and freedom of movement, which is a critical to their psychological needs. Additionally, from a practical perspective, this dependency is suboptimal, as friends and family may not always be available or fluent in ASL themselves. In emergency situations, where immediate and clear communication in ASL is critical, lacking equitable access in communication can become even more distressing and dangerous, further highlighting the need for more inclusive and accessible options in everyday life.

Communication deficits result in disempowerment, which can have lasting psychological effects. The inability to engage with others on an equal footing can erode self-esteem and lead to social anxiety, as they may avoid situations where communication barriers are likely [26]. While voice-to-text technology and lip-reading are sometimes available, they fall short of providing the emotional connection and comfort that comes from conversing in one's natural language. These methods are often cumbersome, inaccurate, and still exclude the the cultural connection embedded in ASL, leaving Deaf individuals feeling disconnected. The constant struggle to be understood can contribute to chronic stress, depression, and even isolation [11]. Thus, the communication gap not only limits access to information and services

but also exacerbates loneliness, perpetuating a cycle of exclusion and marginalization. Addressing this issue is essential for both the emotional well-being and mental health of the D/HH community[26].

This work investigated the usage of Sign Language Technology to begin to remedy this issue. In particular, we developed a bi-directional communication system and examined the feasibility of using Sign Language Technology to provide a bi-directional communication system. We additionally found acceptable standards through by conducting four user tests. The implications for such a system are incalculable; if deployed properly, such technology could change the lives of many tens of millions across the world. This research demonstrates that such a tool is feasible, possible, and here. However, technology such as this must be deployed carefully as to not cause inadvertent harm.

While our system demonstrates encouraging results, it is clear that the successful deployment of such technology requires ongoing refinement and careful integration into real-world contexts. The accuracy and efficiency achieved in our studies mark only the beginning of what is possible. The hybrid AI-human model we tested offers a promising direction for further improving communication outcomes, but its long-term effectiveness in diverse, high-stakes environments remains to be fully understood. Furthermore, as our usability tests show, users desire a system that not only performs well but also respects their communication preferences and cultural nuances, further underscoring the need for personalized, adaptive technology.

Given these findings, it is imperative that future research continues to focus on enhancing the real-world applicability and robustness of AI-powered sign language systems. Longitudinal studies will be crucial in assessing the sustained impact of these tools on daily communication. Moreover, interdisciplinary collaboration with the D/HH community will ensure that the technology evolves in a way that is both technically sound and socially responsible. With this in mind, we now turn to our recommendations for future research and responsible deployment.

We advocate first for continued research in this field, as currently work on tangible solutions to communication barriers are scarce. Through research such as this, we can continue to push the status quo. Currently research into bi-directional communication systems remains thin, both within the machine learning domain and HCI domain. This study demonstrates that there is an active pressing need for this kind of technology for tens of millions of D/HH individuals across the globe.

We advocate second for the involvement of relevant stakeholders within research. While it is tempting to only ask experts and non-profits in the Deaf space for their guidance, or to only involve the Deaf community in the final step of usability testing (or not to involve them at all), it is critical to engage with the D/HH community through the inception, design, development, evaluation, and deployment of this technology. It is additionally critical to involve members from across the spectrum of the D/HH community, ensuring to engage members of all socioeconomic classes, demographics, communication modalities, and signing styles.

We advocate third for responsible deployment of this technology; while this technology shows great promise, it is important not to inadvertently cause harm through irresponsible deployment. It is critical to safeguard privacy and security during deployment, and not deploy this technology into places where it is not ready or wanted. Additionally, when deploying to high-stakes environments, it is crucial to have fallback integrated, such as an interpreter fallback to transfer to an interpreter when repeated errors are detected.

Significant work remains to be done. This study focused only on the short-term effects of deploying a single system for ASL. While technically, transfer learning to other sign languages should be relatively straightforward, usability concerns persist, particularly regarding user interaction with interfaces in regions where literacy rates may be lower than those for English speakers.

Additionally, while this study established an acceptable bar for usability, it remains unclear whether this threshold represents the true minimum for widespread adoption. There is still much work to be done to fully realize the hybrid interpretation model. In future efforts, we plan to implement this hybrid model and explore the feasibility of removing the "start" and "stop" buttons to enable continuous streaming sign language recognition, and other sign languages. This step will be crucial in improving user experience and advancing the capabilities of real-time sign language interpretation.

7 Acknowledgements

This study was supported through grants made by the National Science Foundation (NSF).

References

[1] Wenda Alifulloh, Dadang Juandi, Aan Hasanah, Dian Usdiyana, and Humam Nuralam. 2024. Research Trends of Mathematics Learning for Deaf Junior High School Students in Indonesia: A Systematic Literature Review. *The Eurasia Proceedings of Educational and Social Sciences* (2024), 24–33.

[2] Maryam Aziz and Achraf Othman. 2023. Evolution and Trends in Sign Language Avatar Systems: Unveiling a 40-Year Journey via Systematic Review. *Multimodal Technologies and Interaction* 7, 10 (2023), 97.

[3] Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. 2024. Neural Sign Actors: A diffusion model for 3D sign language production from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1985–1995.

[4] National Deaf Center. 2024. *Deaf Awareness.* https://nationaldeafcenter.org/resources/deaf-awareness/

[5] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. 2019. Everybody dance now. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5933–5942.

[6] Xingyu Chen, Baoyuan Wang, and Heung-Yeung Shum. 2023. Hand avatar: Free-pose hand animation and rendering from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8683–8693.

[7] Federal Communication Commission. 2002. *Video Relay Service.* https://www.fcc.gov/consumers/guides/video-relay-services

[8] Federal Communication Commission. 2014. *Disability Advisory Committee.* https://www.fcc.gov/disability-advisory-committee

[9] Maartje De Meulder and Nienke Sijm. 2024. "I feel a bit more of a conduit now": Sign language interpreters coping and adapting during the COVID-19 pandemic and beyond. *Interpreting and Society* 4, 1 (2024), 3–25.

[10] Vernandi Dyzel, Rony Oosterom-Calo, Mijkje Worm, and Paula S Sterkenburg. 2020. Assistive technology to promote communication and social interaction for people with deafblindness: a systematic review. In *Frontiers in Education*, Vol. 5. Frontiers Media SA, 578389.

[11] Fellinger. [n. d.]. Mental health of deaf people - the lancet. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(11)61143-4/fulltext

[12] Center for Research on Disability. 2022. *Civilians Living in the Community for the United States and States - Hearing Disability: 2022.* https://www.researchondisability.org/ADSC/build-your-own-statistics

[13] Foster and MacLeod. [n. d.]. https://www.researchgate.net/publication/10612960_Deaf_people_at_work_Assessment_of_communication_among_deaf_and_hearing_persons_in_work_settings

[14] Handtalk. 2012. *Discover the largest Sign Language translation platform in the world.* https://www.handtalk.me/en/

[15] Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. 2017. Avatar digitization from a single image for real-time rendering. *ACM Transactions on Graphics (ToG)* 36, 6 (2017), 1–14.

[16] Janis Kritzinger, Marguerite Schneider, Leslie Swartz, and Stine Hellum Braathen. 2014. "I just answer 'yes' to everything they say": Access to health care for deaf people in Worcester, South Africa and the politics of exclusion. *Patient Education and Counseling* 94, 3 (2014), 379–383. https://doi.org/10.1016/j.pec.2013.12.006

[17] Kit Yung Lam, Liang Yang, Ahmad Alhilal, Lik-Hang Lee, Gareth Tyson, and Pan Hui. 2022. Human-avatar interaction in metaverse: Framework for full-body interaction. In *Proceedings of the 4th ACM International Conference on Multimedia in Asia*. 1–7.

[18] Amy Lederberg. 2022. Special Education Research and Development Center on Reading Instruction for Deaf and Hard of Hearing Students. *https://ies.ed.gov/ncser/RandD/details.asp?ID=1325* (2022).

[19] Carman KM Lee, Kam KH Ng, Chun-Hsien Chen, Henry CW Lau, Sui Ying Chung, and Tiffany Tsoi. 2021. American sign language recognition and training method with recurrent neural network. *Expert Systems with Applications* 167 (2021), 114403.

[20] Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in Bleu: Reevaluating the evaluation of Automatic Machine Translation Evaluation Metrics. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020). https://doi.org/10.18653/v1/2020.acl-main.448

[21] Brenda Nicodemus and Karen Emmorey. 2015. Directionality in ASL-English interpreting: Accuracy and articulation quality in L1 and L2. *Interpreting* 17, 2 (2015), 145–166.

[22] Fajar Ainun Nugroho and Alies Poetri Lintangsari. 2022. Deaf Students' Challenges in Learning English: A Literature Review. *Indonesian Journal of Disability Studies* 9, 2 (2022), 217–224.

[23] Maria Papatsimouli, Panos Sarigiannidis, and George F Fragulis. 2023. A survey of advancements in real-time sign language translators: integration with IoT technology. *Technologies* 11, 4 (2023), 83.

[24] Peter V Paul and Peixuan Yan. 2023. The effects of American Sign Language on English reading proficiency. *American Annals of the Deaf* 167, 5 (2023), 745–760.

[25] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10975–10985.

[26] Person. 2024. Hearing loss and anxiety: Why it happens and how to Cope. https://www.ncoa.org/adviser/hearing-aids/hearing-loss-anxiety/#:~:text=Social%20impact:%20Communication%20difficulties%20can,1%2C%202000.

[27] Varshini Prakash and BK Tripathy. 2020. Recent advancements in automatic sign language recognition (SLR). In *Computational Intelligence for Human Action Recognition*. Chapman and Hall/CRC, 1–24.

[28] Lorna Quandt. 2020. Teaching ASL signs using signing avatars and immersive learning in virtual reality. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–4.

[29] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. 2020. Hand sign language recognition using multi-view hand skeleton. *Expert Systems with Applications* 150 (2020), 113336.

[30] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. 2022. Real-time isolated hand sign language recognition using deep networks and SVD. *Journal of Ambient Intelligence and Humanized Computing* 13, 1 (2022), 591–611.

[31] Amit Shankar and Biplab Datta. 2020. Measuring e-service quality: a review of literature. *International Journal of Services Technology and Management* 26, 1 (2020), 77–100.

[32] Shikhar Sharma and Krishan Kumar. 2021. ASL-3DCNN: American sign language recognition technique using 3-D convolutional neural networks. *Multimedia Tools and Applications* 80, 17 (2021), 26319–26331.

[33] Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Julien Valentin, Jie Song, and Otmar Hilliges. 2023. X-avatar: Expressive human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16911–16921.

[34] Kristin Snoddon. 2016. Whose ASL counts? linguistic prescriptivism and challenges in the context of parent sign language curriculum development. *International Journal of Bilingual Education and Bilingualism* 21, 8 (Sep 2016), 1004–1015. https://doi.org/10.1080/13670050.2016.1228599

[35] William C Stokoe Jr. 2005. Sign language structure: An outline of the visual communication systems of the American deaf. *Journal of deaf studies and deaf education* 10, 1 (2005), 3–37.

[36] Fatma M Talaat, Walid El-Shafai, Naglaa F Soliman, Abeer D Algarni, Fathi E Abd El-Samie, and Ali I Siam. 2024. Real-time Arabic avatar for deaf-mute communication enabled by deep learning sign language translation. *Computers and Electrical Engineering* 119 (2024), 109475.

[37] DEAF SERVICES UNLIMITED. 2023. *ASL Interpreter Shortage and Accessibility in Higher Education.* https://deafservicesunlimited.com/asl-interpreter-shortage-and-accessibility-in-higher-education/

[38] Prokopia Vlachogianni and Nikolaos Tselios. 2022. Perceived usability evaluation of educational technology using the System Usability Scale (SUS): A systematic review. *Journal of Research on Technology in Education* 54, 3 (2022), 392–409.

[39] Rosalee Wolfe, John C McDonald, Thomas Hanke, Sarah Ebling, Davy Van Landuyt, Frankie Picron, Verena Krausneker, Eleni Efthimiou, Evita Fotinea, and Annelies Braffort. 2022. Sign language avatars: a question of representation. *Information* 13, 4 (2022), 206.

[40] Fei Xu, Lipisha Chaudhary, Lu Dong, Srirangaraj Setlur, Venu Govindaraju, and Ifeoma Nwogu. 2024. A Comparative Study of Video-Based Human Representations for American Sign Language Alphabet Generation. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–6.

[41] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. 2024. Vasa-1: Lifelike audio-driven talking faces generated in real time. *arXiv preprint arXiv:2404.10667* (2024).

[42] Andrea Xue Jin Yet, Vinuri Hapuhinne, Weilyn Eu, Elizabeth Yie-Chuen Chong, and Uma Devi Palanisamy. 2022. Communication methods between physicians and Deaf patients: a scoping review. *Patient Education and Counseling* 105, 9 (2022), 2841–2849.