Bidirectional Language Agnostic Framework for Sign Language Production and Recognition

By

Mateen Ahmed

(Registration No: 00000329700)

Department of Robotics & Artificial Intelligence

School of Mechanical & Manufacturing Engineering

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

(2024)

Bidirectional Language Agnostic Framework for Sign Language Production and Recognition

By

Mateen Ahmed

(Registration No: 00000329700)

A thesis submitted to the National University of Sciences & Technology, Islamabad,

in partial fulfillment of the requirements for the degree of

Master of Robotics & Intelligent Machine Engineering

Supervisor: Dr. Muhammad Tauseef Nasir

School of Mechanical & Manufacturing Engineering

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

(2024)

NUST Campus Management Solution
Pakistan

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by **Regn No. 00000329700 Mateen Ahmed** of **School of Mechanical & Manufacturing Engineering (SMME)** has been vetted by undersigned, found complete in all respects as per NUST Statues/Regulations, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis titled. **Bidirectional Language Agnostic Framework for Sign Language Production and Recognition**

Signature:

Name (Supervisor): Muhammad Tauseef Nasir

Date: 28 - Aug - 2024

Signature (HOD):

Date: 28 - Aug - 2024

Signature (DEAN):

Date: 28 - Aug - 2024

Form TH-4

National University of Sciences & Technology (NUST)

MASTER'S THESIS WORK

We hereby recommend that the dissertation prepared under our supervision by: Mateen Ahmed (00000329700)
Titled: Bidirectional Language Agnostic Framework for Sign Language Production and Recognition be accepted in partial fulfillment of the requirements for the award of MS in Robotics & Intelligent Machine Engineering degree.

Examination Committee Members

1. Name: Waqas Khalid Signature:

2. Name: Muhammad Asim Waris Signature:

Supervisor: Muhammad Tauseef Nasir Signature:

Date: 28 - Aug - 2024

Head of Department

28 - Aug - 2024

Date

COUNTERSINGED

28 - Aug - 2024

Date

Dean/Principal

CERTIFICATE OF APPROVAL

This is to certify that the research work presented in this thesis, entitled "Bidirectional Language Agnostic Framework for Sign Language Production and Generation" was conducted by Mr. Mateen Ahmed under the supervision of Dr. Muhammad Tauseef Nasir.

No part of this thesis has been submitted anywhere else for any other degree. This thesis is submitted to the Department of Robotics & Artificial Intelligence in partial fulfillment of the requirements for the degree of Master of Science in the Field of Robotics and Intelligent Machine Engineering at School of Mechanical and Manufacturing Engineering, National University of Sciences and Technology, Islamabad.

Student Name: Mateen Ahmed   Signature: ______________.

Examination Committee:

a) Internal Examiner 1: Dr. Muhammad Asim Waris   Signature: ______________.

(Associate Professor & HoD BMES)

……………………………….

b) External Examiner 2: Dr. Waqas Khalid   Signature: ______________.

(Assistant Professor)

………………………………

Supervisor Name: Dr. Muhammad Tauseef Nasir   Signature: ______________.

Name of Dean/HOD: Dr. Kunwar Faraz   Signature: ______________.

AUTHOR DECLARATION

I Mateen Ahmed hereby state that my MS thesis titled *"Bidirectional Language Agnostic Framework for Sign language Production and Recognition"* is my own work & has not been submitted previously by me for taking any degree from National University of Sciences & Technology, Islamabad or anywhere else in the country/ world.

At any time if my statement is found to be incorrect even after I graduate, the university has the right to withdraw my MS degree.

matt

Mateen Ahmed

2020-NUST-MS-R&AI-329700

PLAGIARISM UNDERTAKING

I solemnly declare that research work presented in the thesis titled **"Bidirectional Language Agnostic Framework for Sign language Production and Recognition"** is solely my research work with no significant contribution from any other person. Small contribution/ help wherever taken has been duly acknowledged and that complete thesis has been written by me. I understand the zero tolerance policy of the HEC and National University of Sciences and Technology (NUST), Islamabad towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited. I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS degree, the University reserves the rights to withdraw/revoke my MS degree and that HEC and NUST, Islamabad has the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized thesis.

Signature of Student: ~matt-

Name of Student: Mateen Ahmed

DEDICATION

Dedicated to the deaf, blind & other handicapped communities in Pakistan & all over the world. May this miniscule endeavor pave way for general betterment, widespread facilitation & social inclusion of these communities in the near future.

ACKNOWLEDGEMENTS

First & foremost, all gratitude & praise belongs to Almighty Allah, the creator, sustainer, protector, & the Lord of the Worlds. May the fruits of this labor be practically materialized in the near future for those in need for social inclusion & reduction in communication barrier. Secondly, my grandparents & parents for their prayers; for inculcating important values like commitment & hard work as well as all my teachers, mentors & professors especially Dr. Hasan Sajid, Dr. Muhammad Jawad Khan & Dr. Muhammad Tauseef Nasir, whose words of wisdom may not have been in vain after all.

TABLE OF CONTENTS

LIST OF TABLES

Page No.

LIST OF FIGURES

Page No.

LIST OF SYMBOLS, ABBREVIATIONS & ACRONYMS

AI	Artificial Intelligence
ASL	American SL
DL	Deep Learning
GL	Gloss
LLM	Large Language Model
NLP	Natural Language Processing
NMFs	Non-Manual Features
SL	SL
SLP	SL Production

ABSTRACT

This thesis explores the development of an automated SL production system using cutting-edge advancements in natural language processing (NLP) & computer vision. Motivated by the growing need for inclusive communication solutions, especially for the deaf & hard-of-hearing community, the project focuses on generating SL poses from input text. The proposed system leverages a series of interconnected processes, including semantic & grammar correction, translation, word validation & substitution, pose generation, & pose stitching. By ensuring language agnosticism & adaptability across multiple languages, the system aims to bridge communication gaps for deaf individuals in various social, legal, & corporate settings. This research is grounded in an extensive review of the current methodologies in SL recognition & production, highlighting the limitations & potential improvements in existing systems. The final evaluation of the system demonstrates promising results in producing accurate & comprehensible SL videos, contributing to the ongoing efforts to promote accessibility & inclusivity for the deaf community.

Keywords: SL Production, Computer Vision, Large Language Models, Multilingual Translation, Accessibility & Inclusivity

CHAPTER 1: INTRODUCTION

This chapter will briefly describe the project overview its goals & objectives. It gives a comprehensive detail of the domain & challenges that this project has. Rapid advancements in generative AI, especially in the domain of large language models (LLMs), have opened up new possibilities for applications in various fields. Recognizing this trend & the growing capabilities of LLMs, the author was motivated to utilize such technologies in an impactful project. This project was driven by two main motivations: leveraging the capabilities of LLMs & designing a system for SL production.

1.1 Motivation

Ensuring the fundamental right to communication for all individuals is a cornerstone of a just society. To preserve the right to communication to every person is the base of a fair society. It crucial to deal with the problems of communication that people, who are vocally impaired & depend on SL, are facing. Although, sign language is the most unique or expressive method of communication for the deaf community [1], but at the same time it also leads to a significant limitation that hearing individuals are unable to understand, which in turn, causes a deterioration in the expressions of feelings and socialization opportunities. This exclusion comes from the fact that hearing persons are not able to understand sign languages, this than results in a significant communication gap that should be eliminated. For a society to function as a whole it is important to address the communication barriers faced by those who are vocally disabled and use sign language. Although sign language provides a rich way of interaction within the deaf community [1], it often limits interactions with the broader hearing world, hindering the sharing of thoughts and emotions.

1.2 Significance

Communication is the bedrock of human connection, it enables the exchange of ideas, and experiences. Although speech serves as a primary mode of communication for many people, it is important to recognize & accommodate the unique communication needs of deaf people. Sign language, a visually expressive but nuanced language, plays a key part in easing communication for individuals who cannot speak. However, the lack of widespread understanding of sign language among hearing individuals creates a big barrier to social inclusion, this also limits the full participation of deaf individuals in societies [2].

Modern advancements in AI provide promising solutions for overcoming such communication barriers as evident from [4] . In particular the machine learning, and deep learning techniques can be leveraged to facilitate the production of SL, in a society this will enable seamless interaction between deaf & hearing individuals. This recent technology can also be harnessed to translate spoken or written language into sign language.

1.3 Importance of SL

SL, a visually-rich & expressive language, serves as the primary mode of communication for deaf individuals around the globe. Various approaches have been proposed for generating SL from text or speech, utilizing computer vision techniques & linguistic models. As sign language encompasses a diverse range of hand postures, movements, facial expressions, & body language, forming a complex & nuanced system of communication. While sign language is widely used within the deaf community, its limited understanding among hearing individuals creates a communication gap that stops interaction among members of society in accordance with the author's experience. Automatic systems for sign language production have emerged as a potential solution to solve this problem, enabling the translation of spoken or written language into Sign lang.

Research on automatic sign language production can be categorized into isolated sign language production and continuous sign language production. These methods help capture the structure of sign language which is quiet complex, including handshapes, movements of body, facial expressions, and body language, to produce and accurate description of ideas being conveyed.

In conclusion, development of an robust or reliable SL production systems is crucial for promoting social-inclusion & providing means for deaf individuals to communicate effectively in their own language. While challenges remain, ongoing research & development efforts are clearing the way for innovative solutions that can bridge the communication gap between deaf & hearing individuals.

1.4 Purpose

The core objective of our project is to help deaf individuals who use SL to communicate effectively with those who do not understand it. Furthermore, our project will focus specifically on language agnosticism – development of a system that can cater for a wide range of languages. By developing a SL production system, this project demonstrates a commitment to addressing the specific needs of the SL community & facilitating their full participation in society.

1.5 Challenges

- Developing a system that can capture all the features of SL is challenging due to the extensive vocabulary used in daily communication, which typically consists of thousands of signs.
- Despite the large vocabulary, certain words, such as personal names, might not be included; in these cases, it may be necessary to represent those words with signs at the character level.
- The movements of the body, hands, & head are crucial in accurately recognizing signs.
- Even minor differences between two signs can lead to incorrect interpretations if not properly identified.
- A single SL term may have multiple equivalents in natural language. For example, depending on the context, the sign for "want" could also be understood as "hungry." Additionally, words sharing a lemma with a noun or verb might be represented by the same sign. These subtleties are often not well-represented in existing small-scale databases.

- Modeling the deformable, articulated structure of the hand, along with self-occlusion, presents significant challenges.
- Many applications require tracking hands in cluttered environments & under varying, poorly controlled lighting conditions.
- Simultaneously tracking both hands requires solving the temporal matching (data association) problem & managing temporary occlusions when one hand blocks the view of the other.
- Hands can also be obscured by clothing, with occlusion ranging from partial (e.g., long sleeves) to full (e.g., gloves).

The challenges associated with SL production are multifaceted. The vast vocabulary of SL, the subtle variations in handshapes & movements that can alter meaning, & the integration of facial expressions & body language all contribute to the complexity of the task.

1.6 Objective

The main purpose of this study is to provide methods & to create an automated system that can produce SL pose from input text & in doing so contribute towards the individuals that suffer from hearing loss & promote their social inclusion.

1.7 Problem Statement:

Approximately 430 million people, accounting for over 5 percent of the global population, experience hearing impairment. Learning SL can be challenging & time-consuming, leading to a communication gap between hearing individuals & those who are hard of hearing. To facilitate interaction with deaf individuals, natural language must often be translated into SL by an interpreter. Therefore, developing a SL production algorithm is essential. Such a tool would significantly benefit the deaf community & help bridge the communication gap.

1.8 Areas of Application

Major areas of the application for SL production systems are the following:

- Accessibility & Inclusivity: Improves communication for the deaf & hard-of-hearing community across various settings.
- Education: Assists in teaching SL & provides educational content in SL for deaf students.
- Healthcare: Facilitates communication between healthcare providers & deaf patients.
- Media & Entertainment: Enables content creators to include SL in videos, movies, & TV shows for broader accessibility.
- Public Announcements: Ensures emergency alerts & public service announcements are accessible in SL.
- Customer Service: Enhances customer support for deaf customers through SL video responses.
- Legal & Judicial Settings: Improves accessibility in legal proceedings & courtrooms by providing SL interpretation.
- Corporate Communication: Supports inclusive workplace communication by offering training & HR materials in SL.

1.9 Deliverables

- End-to-end system for SL production.
- Language agnosticism of developed system.
- Working modalities: Text & Pose

1.10 Thesis Organization

The thesis is organized as follows: Chapter 2 details the previous research carried out for SL production, its significance & different strategies achieved so far. The chapter also identifies the research gap. Chapter 3 discusses the proposed methodology & the implementation details, including the state-of-the-art model pipeline used for SL production. Chapter 4 discusses the results achieved after implementing the proposed methodology, while Chapter 5 gives the conclusions drawn based on those results. It also gives future direction as to how this work can be enhanced.

CHAPTER 2: LITERATURE REVIEW

Over the past two decades, an extensive body of literature has emerged focusing on various global & local SLs, as well as the different approaches used to address SL recognition through image-based models. A substantial part of this research has concentrated on American SL (ASL), given its widespread use & popularity within the deaf communities in the United States & other countries. Only in recent years have vision & image-based systems for the regional SLs of other countries begun to gain attention in academic circles.

SLs, like natural spoken languages, exist & develop independently from one another. The SLs are the primary mode of communication for deaf individuals, which constitutes hands' movements, as well as face expressions & body cues. Unlike spoken languages, which are perceived in oral-auditory physical modality, SLs require the manual-visual physical modality for proper understanding & comprehension of inter-personal communication [1]. For example, ASL has developed independently from English language spoken in the United States, having more than 22 million (10% of the American population) deaf individuals [2-3], out of which many are unable to properly communicate or function in their daily lives, despite rise in technology & related applications [4-5]. It is estimated that there are more than 6,909 spoken languages around the world, while their SL counterparts amount to only 138 [6]. Although widely being used as far back as the early eighteenth century, but due to lack of proper assessment tools, SLs were not fully understood & considered a universally generalized & limited from of pantomime up till the 1950s[7]. However, structural analysis of SL by Tervoort [8] shed light on the individual & separately identify SLs from the spoken languages, unlike previously assumed. His findings revealed that SLs followed properties akin of an independent linguistic system in the different modality, as each word & its corresponding gesture had a constant form & meaning, derived independently from the spoken language [8]. Similarly, Stokoe [9] ground-breaking work on ASL allowed academia to take linguistic interest in SLs in the manner like spoken languages. He developed a model for describing ASL by outlining its phonological parameters in the form of variations in hand shapes, arm movements, body cues & facial expressions [9]. As a result, till the 1980s & 1990s, adequate attention was being given to promote SLs & their various facets in order to firmly establish the fact that SLs were separate languages with their unique linguistic & lexical identities & forms [1].

SLs are learned & acquired by children in the same manner as spoken languages, while being prone to make similar mistakes such as phonological substitution, omission & over-generalization [10–12]. Similarly, signing language can be essentially 'created' with the induction of deaf children within domestic, or educational settings; thereby giving rise to a unique SL (with proper grammatical structure), along with its communal values & 'culture', as is reported for a deaf school in Managua, Nicaragua [13]. The knowledge regarding SLs within the general communities is very limited & fraught with misconceptions that ultimately gives rise to negative stereotypes & sense of alienation of the deaf individuals [14]. Some of the general misconceptions include the following [3, 15]:

1. Spoken & gesture languages are related in terms of their structure & syntax.
2. Signers around the world can understand one another.
3. All global SLs have evolved in similar manner[3], [15].

Although, there has been numerous reporting of indigenous communities within Desa Kolok (Bali), Negev desert (Israel) & Adamorobe (Ghana), where genetically transmitted deafness & high deaf population allowed the local hearing population to achieve fluency in local SLs [16,17]. As a result, deafness & local deaf individuals are not stigmatized within such communities. However, stigmatization & marginalization within contemporary societies is generally promoted with lack of adequate tools to assist communication between local deaf community & hearing population. Similarly, within Pakistan, disability in general & deafness in particular is normally concealed & highly stigmatized, causing social, psychological & emotional issues for the deaf individuals [18].

2.1 Similarities & Variations among SL Systems

Even though this field has not developed as much as speech recognition & is considered in its early stages by Cooper et al. [19]. Before doing deeper into the different techniques for SL Translation being used at present, it is important to understand the different characteristics of SLs, which can be later used to differentiate & recognize different signs. Vogler & Metaxas [20] shed light on phoneme & using the extension of Movement-Hold model by Liddell [21] for ASL recognition, which essentially splits each gesture into a sequence of phonemes or a sequence of movements & holds. Following are some of the unique features of SL identified by Cooper et al. [19]:

- Classifiers: Hand shapes used to represent a particular class of objects
- Positional Signs: When signs are acting on a specific part of the body descriptively, such as an injury or tattoo.
- Non-manual Features: Mouth expressions & physical postures conjunction with hand gestures can provide information about the meaning of sentence or context of discussion, such as lip spacing, or shape, & eyebrow position.
- Directional Verbs: The direction of motion during performing signs indicates the direction of verbs between the signer & referents. Examples include 'give phone' or 'take money'.
- Body shifts: These are used to represent role-shifting when relating a dialogue by twisting shoulders or changing gaze.
- Adverbs modifying verbs: There will not be two different signs for phrases such as 'running quickly' or 'running slowly'; rather they would modify signs for running by either performing sign faster or slower respectively.
- Finger spellings: In case of not knowing the sign for particular word, the signer can manually spell the word explicitly by finger spelling.

2.2 Related Works

Numerous studies have investigated the application of Transformers [22] & Spatial-Temporal Graph Convolutional Networks [23] in analyzing spatial-temporal relationships within non-Euclidean datasets. Traditional Transformers operate on fully connected graphs, where all token connections are represented, potentially resulting in suboptimal performance when the graph topology is not incorporated into the node features. In contrast, Spatial-Temporal Graph Convolutional Networks integrate both spatial & temporal relationships from the data, introducing higher-level semantics that can effectively complement Transformer models.

The integration of graph structures with Transformer models has been applied across various domains. For instance, Guo et al. [24] introduced a self-attention-based graph neural network for traffic prediction, utilizing self-attention to capture temporal dynamics & a graph convolution module to model spatial correlations. In the field of skeleton-based human action recognition, Plizzari et al. [25] used a self-attention mechanism to capture dependencies between joints & employed a two-stream approach to conditionally model the natural structure of the human body. Additionally, Dwivedi et al. [26] extended the Transformer architecture to support arbitrary graphs, incorporating both node & edge feature representations.

The production of SL has emerged as a key challenge in neural machine translation & has garnered considerable attention in recent years [27], [28], [29], [30]. The Transformer model, renowned for its self-attention mechanism that operates without convolution, has demonstrated exceptional success in natural language processing. In this context, Saunders et al. [33] introduced the first Transformer-based model for SL production, which learns to map spoken language sentences to sign pose sequences in an end-to-end framework. However, many of these approaches translate sign pose sequences into Euclidean data, neglecting the inherent structure, semantics, & other crucial characteristics of skeletal data.

To address this limitation, Saunders et al. [34] introduced a spatial-temporal skeletal graph attention layer that incorporates a hierarchical body inductive bias into the self-attention mechanism. Additionally, Huang et al. [35] developed spatial-temporal graph convolution layers for the pose generator, effectively capturing both intra-frame & inter-frame information in SL videos. However, despite these advancements, many methods fail to account for the varying importance of each joint in conveying gesture meaning. Critical factors such as motion relationships & the amplitude of actions play a significant role in determining the meaning of SL. To better represent non-Euclidean data, the author proposes a novel graph partitioning strategy that separately models the upper limb & hand regions.

In human mesh reconstruction, a model is used to create a 3D representation of the skin. The Skinned Multi-Person Linear model [36] is commonly used to describe important features of the human body, such as different body shapes & natural poses. This model uses a skeleton structure to control the shape of the mesh, with 6,890 points, 13,776 triangles, & 24 joints. However, while it accurately reconstructs the body, it lacks detail in the hands.

To fix this, the Hand Model with Articulated & Non-Rigid Deformations [8] offers a solution by creating a direct link between hand poses & adjustments to the mesh for better accuracy. Similarly, the Faces Learned with an Articulated Model & Expressions model [31] focuses on detailed head reconstruction, capturing head rotation & modeling the neck area. A more advanced model combines all of these—body, head, & hand—resulting in a complete 3D representation that includes body position, hand movements, & facial expressions.

2.3 SLP Taxonomy

In this section, we present a structured overview of the fundamental concepts in deep learning applied to SL Production. We organize recent studies into distinct categories &

analyze each one comprehensively. The upcoming subsections will explore different types of input, datasets, use cases, & the various models discussed in current research.

2.3.1 SL Production Modalities

In SL Production, two primary input modalities are typically used: visual & linguistic. The visual modality encompasses video & images, while the linguistic modality involves natural language. Processing these modalities requires computer vision & natural language techniques.

2.3.2 Visual Modality

In SL Production models, visual input generally falls into two types: RGB images or videos & skeleton data. RGB images & videos provide detailed content but are more complex due to their high dimensionality. On the other hand, skeleton data simplifies the model & speeds up processing by reducing dimensionality.

For RGB images, each frame can show a single letter or digit, & deep learning techniques are used to extract important features. Convolutional Neural Networks are particularly effective for this task [37]. Generative models, like Generative Adversarial Networks, use Convolutional Neural Networks to create sign images or videos. However, RGB videos are more complex because they include the temporal dimension, making them harder to process than single images. Most SL Production models use RGB videos [38, 39, 40], which can show individual signs or sequences of signs forming sentences. To handle both static & dynamic visual data, Generative Adversarial Networks & Long Short-Term Memory networks are often used. Although these models have made significant progress, more development is needed to create more lifelike sign images & videos, improving communication tools for the deaf community.

2.3.3 Linguistic Modality

Text input remains the predominant linguistic modality in SL Production. A variety of sophisticated models are utilized to handle textual input [41, 42]. Although text processing is generally less intricate compared to image or video processing, it presents its own set of complexities, particularly in the realm of translation. Neural Machine Translation (NMT) models are extensively employed for the processing of text input. Additionally, Sequence-to-Sequence models, such as Recurrent Neural Networks (RNNs), have demonstrated their efficacy across a range of applications.

Despite these advancements, translation tasks continue to face persistent challenges, such as domain adaptation, due to the variability in lexical styles, translations, & meanings across different languages. Addressing these issues often necessitates a focus on specific domains. Transfer learning—where a translation model is initially trained on general data & subsequently fine-tuned on domain-specific data—serves as a prevalent strategy to tackle this problem. Moreover, the volume of training data is pivotal, as deep learning models typically require extensive datasets to achieve robust generalization. Machine translation systems also grapple with rare & previously unseen words, a challenge that can be alleviated through byte-pair encoding techniques, including stemming & compound-splitting. Long sentences pose additional difficulties, though attention mechanisms can offer partial relief, particularly for shorter sentences. Furthermore, word alignment issues become increasingly significant in reverse translation, where the process involves translating text back from the target language to the source language.

2.4 Datasets

Although numerous extensive annotated datasets are available for SL recognition, there remains a paucity of large-scale datasets specifically curated for SL Production. Two prominent datasets frequently employed in SL translation are RWTH-Phoenix-2014T [43] & How2Sign [44]. The RWTH-Phoenix-2014T dataset encompasses sentences in German SL, rendering it instrumental for text-to-SL translation. This dataset is an extension of the continuous SL recognition corpus, PHOENIX-2014 [45], & comprises 8,257 sequences performed by 9 signers, incorporating 1,066 sign glosses & 2,887 spoken language terms. It also provides gloss annotations aligned with spoken language sentences. Figure 2 illustrates an overview of these datasets.

The How2Sign [40] dataset, a more contemporary multi-modal dataset, is tailored for speech-to-SL translation. It encompasses 38,611 sequences & 4,000 vocabularies executed by 10 signers. In a manner similar to RWTH-Phoenix-2014T, it includes sign gloss annotations.

Despite the existence of these evaluative benchmarks, both RWTH-Phoenix-Weather 2014T & How2Sign fall short in facilitating the generalization of SL Production models. Furthermore, these datasets are limited to German & American SL sentences. To develop an efficient application for enhancing communication between the deaf & hearing communities, there is a critical need for new, extensive datasets that offer substantial variety & diversity across multiple SLs.

2.5 Proposed Models

In this section, we review recent advancements in SLP categorized into five key approaches: Retargeting, NMT, Motion Graphs, Diffusion based Image or Video Generation, & Other Models. Table 1 summarizes the models for SLP.

To overcome communication obstacles between hearing & hearing-impaired individuals, sign avatars are being examined as an economical substitute for live interpreters. Sign avatars utilize three-dimensional animated models to illustrate SL gestures, providing an efficient method to present signed dialogues without relying on actual video footage. These avatars can portray intricate movements, encompassing fingers, hands, facial expressions, & body motions, & can be programmed for a variety of SLs. Recent advancements in computer graphics have enabled the production of high-quality animations with smooth transitions between signs.

Sign avatars may be produced from motion capture data or parametrized glosses. Noteworthy examples in this field include VisiCast [46], Tessa [47], eSign [48], dicta-sign [49], JASigning [50], & WebSign [51]. These models often require annotated sign videos, such as those encoded using HamNoSys [52] or SigML [53]. However, issues such as unnatural movements, absent non-manual cues (e.g., eye gaze, facial expressions), & the uncanny valley phenomenon have restricted their acceptance in the deaf community [54]. Recent efforts are focusing on incorporating non-manual information to enhance avatar authanticity [55, 56].

Motion capture-based avatars, like Sign3D by MocapLab [57], offer exceptionally realistic outcomes but are constrained by the expense of data collection. These models are not readily scalable & demand expert oversight. To address these constraints, recent innovations in deep learning & graphical methodologies, including Motion Graphs, are being utilized [58].

2.5.1 NMT Approaches

Machine Translation (MT) has evolved significantly since its inception in the 1960s, with methods ranging from rule-based to statistical & example-based approaches [59]. In SLP, translating spoken language into SL is complex due to mismatches in gloss ordering & quantity between spoken & signed languages.

Neural Machine Translation has become prominent in translating text to sign employing Neural Networks. NMT models, such as those improved by Luong et al. [60], predict word sequences & model entire sentences. Hybrid models combining CNNs & sequence-to-sequence (seq2seq) architectures [61] or 3D Convolutional Neural Networks (3DCNNs) with LSTM-based encoders [62] have shown improved performance. Additionally, approaches like dilated convolutions & Transformers [63,64] are also used.

It is crucial to emphasize that Stoll et al. [58] developed a model that leverages Neural Machine Translation, Generative Adversarial Networks, & motion synthesis techniques to generate signing videos from linguistic sentences with minimal video annotation. Although methods based on Neural Machine Translation have demonstrated significant achievements, challenges such as adapting to different domains, managing data volume, addressing rare words, & ensuring accurate word alignment persist [62]. Figure 3 illustrates the translation outcomes from [40].

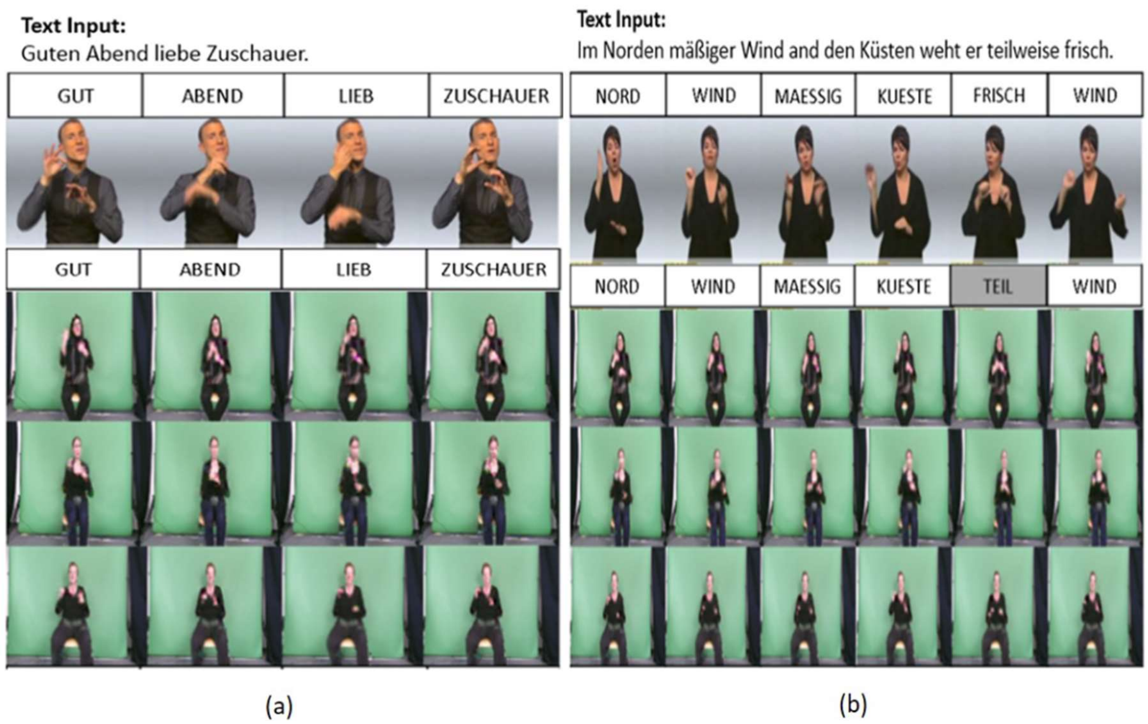

Figure. 1 (a), the German phrase "Guten Abend liebe Zuschauer" is translated, while (b) features "Im Norden mässiger Wind an den Küsten weht er teilweise frisch. The top row displays the GT gloss with image, whereas the row below presents the produced gloss & image. This model combines a Neural Machine Translation network with a GAN to enhance user experience [90].

2.5.2 Motion Graph Approaches

Motion Graph (MG) is a computational graphics method that employs directed graphs derived from motion capture information to produce novel sequences. MG can be combined with Neural Machine Translation-based frameworks for ongoing text-to-pose conversion. Initial endeavors by Kovar et al. [45] concentrated on deriving specific graph traversals & establishing frame intervals. Similarly, Lee et al. [65] proposed a dual-layer framework, utilizing a Markov model & applying clustering techniques.

In addition to the aforementioned, Stoll et al. [58] applied MG to pose data, integrating sign glosses into an MG with transition probabilities facilitated by a Neural Machine Translation decoder. Although MG can produce realistic & modifiable motion, it faces obstacles including data accessibility, scalability issues, & computational demands due to expanding branching factors in the search algorithms.

2.5.3 Conditional Image/Video Generation

Recent developments in advanced machine learning have profoundly influenced the automated creation of images & videos. Key obstacles include maintaining coherence between frames & producing authantic human motion. Machine learning-based frameworks such as conditional Variational Autoencoders & Generative Adversarial Networks are utilized for this purpose.

Combined models that integrate Variational Autoencoders & Generative Adversarial Networks are especially pertinent for producing images & videos of individuals using SL. Innovations such as Van den Oord et al. Pixel Recurrent Neural Networks, Gregor et al. Recurrent Neural Network-based framework, & Kataoka et al. Generative Adversarial Network-attention model have shown enhancements in image synthesis. Despite these improvements, issues like mode collapse, lack of convergence, & stability in training continue to challenge researchers.

2.5.4 Other Models

Several frameworks have advanced SL Production through diverse deep learning methods. For instance, Saunders et al. [39] proposed Progressive Transformers for generating continuous sign sequences from spoken language. Kanis [76] developed a synthesis framework focusing on skeletal model creation using feed-forward & recurrent transformers. Saunders et al. [77] introduced a generative model that combines transformers with Mixture Density Networks for producing photo-realistic sign videos from text. Tornay et al. [78] created a multi-channel evaluation method incorporating hand shape, motion, mouthing, & facial expressions.

Despite these advances, challenges related to model complexity & balancing precision with task complexity remain unresolved. An overview is provided in Table 1.

Table 1. Summary of SL production models.

Year	Ref	Feature	Input	Dataset	Description
2011	[79]	Armature	Frames	ViSiCAST	**Pros:** Focuses on animation content; evaluates using a new metric for comparing Armatures with human signers. **Cons:** Needs to incorporate non-manual features of human signers.
2016	[80]	Armature	Frames	Self Developed	**Pros:** Adds realism to generated images automatically; low computational complexity. **Cons:** Requires accurate shoulder & torso positioning relative to the Armature elbow, rather than the IK end-effector.
2016	[57]	Armature	Frames	Self Developed	**Pros:** High viewer understanding of signing Armatures. **Cons:** Limited to a small set of sign phrases.
2018	[61]	NMT	Frames	PHOENIX-Weather 2014T	**Pros:** Reliable in aligning, recognizing, & translating sign videos. **Cons:** Requires alignment of language glosses & signs.
2018	[62]	NMT	Frames	Self Developed	**Pros:** Reliable in aligning word order with visual content in sentences. **Cons:** Needs to generalize to additional datasets.
2020	[40]	NMT, MG	Text	PHOENIX14T	**Pros:** Effective with minimal gloss & skeletal level annotations for model training. **Cons:** High model complexity.
2020	[39]	Others	Text	PHOENIX14	**Pros:** Reliable to varying lengths of output sign sequences. **Cons:** Model performance can be enhanced by including non-manual information.
2020	[76]	Others	Text	Czech news	**Pros:** Reliable to missing parts of the skeleton. **Cons:** Model output can be enhanced by including mouth expression information.
2020	[77]	Others	Text	PHOENIX14T	**Pros:** Reliable in producing non-manual features.

					Cons: Needs to enhance the realism of generated signs.
2020	[38]	Others	Text	PHOENIX14T	**Pros:** Does not require gloss information. **Cons:** High model complexity.
2020	[81]	Others	Text	PHOENIX14T	**Pros:** Good in producing NMFs. **Cons:** Needs to improve the generated signs.
2021	[86]	YOLO	Text	Self Developed	**Pros:** Used YOLO for sign detection. **Cons**: Can only work on a handful of gestures.
2023	[87]	Diffusion	Text	PHOENIX-Weather 2014T	**Pros**: Can generate sign using diffusion. **Cons**: Poor Performance, most signs generated were incorrect.
2024	[88]	Diffusion	Text	Self Developed	**Pros**: Diffusion-based SLP model, generates dynamic sequences using diffusion. **Cons:** Limited Vocab

CHAPTER 3: METHODOLOGY

The goal of this project is to develop a pipeline for generating SL videos from input sentences using a series of natural language processing (NLP) & computer vision techniques. The pipeline consists of five main stages: Semantic & Grammar Correction, Translation, Word Validation & Substitution, Pose Generation, & Pose Stitching. Below is a detailed description of each stage & the methodologies employed, along with examples to illustrate each process. Figure 2 shows the complete pipeline of our method.

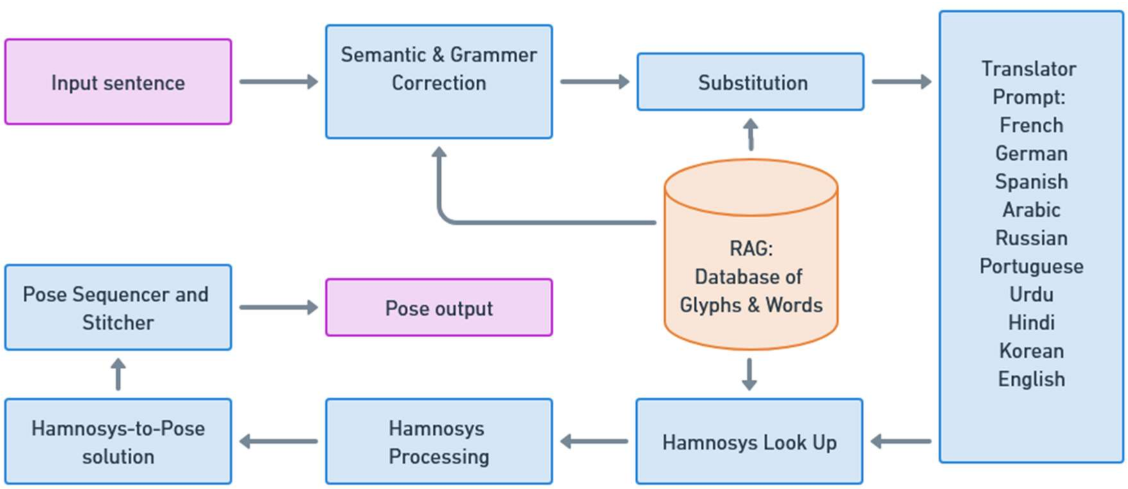

Figure. 2 Pipeline of the Complete Model

3.1 Semantic & Grammar Correction

The prior stage of this pipeline focuses on ensuring that the input sentence is grammatically correct & free of informal language . This has been achieved using the Gemini Nano model, a large language model which performs well for semantic & grammatical corrections. The tasks involved in this stage are discussed below:

- **Spell Checker**: The model corrects any spelling errors & replaces slang with formal equivalents.
    - **Input**: "I dunno, that sounds kinda sus."
    - **Output**: "I don't know, that sounds suspicious."

- **Semantics & Grammatical Analysis**: Here the model analyzes the sentence for structure & semantic coherence, it makes necessary changes to ensure the sentence is grammatically sound.
  - **Input**: "He goes to school yesterday.”
  - **Output**: “He went to school yesterday."
- **Conditional Correction**: Based on the sentences context, the model Will perform conditional corrections to ensure that the sentence maintains its intended Meaning.
  - **Input**: "She don’t get no money.”
  - **Output**: “She doesn't have any money."
- **Appropriateness Checks & Tone Orientation**: This step adjusts the tone of the sentences based on the context. This module is optional.
  - **Input** :"This is a pretty good result."
  - **Output**: "This is a satisfactory result."
- **Stemming & Lemmatization**: Our model also applies stemming & lemmatization to standardize word forms, this is crucial for subsequent processing stages.
  - **Input**: {He is improving his skills by improvements in various areas."
  - **Output**: ‘He improves his skills by making improvements in various areas."

3.1.1 Input Model Query

Input: [Insert the sentence here]

Output: [Provide the corrected version of the sentence here]

Instructions:

1. Spell Check: Carefully examine the input sentence for any spelling errors. Correct all misspelled words to ensure accuracy & adherence to standard language conventions.

2. Gram Mar Correction: Analyze the grmmatical structure of the sentence. Correct any grammatical errors, including issues with subject-verb agreement, verb tenses, punctuation, & sentence fragmens, to enhance clarity & coherence.

3. Formal Language Replacement: Identify & replace any slang, informal expressions with informal & contextually appropriate alternatives. The goal is to enhance the pofessionalism & readability of the sentence without altering its original meaning.

4. Output Requirements: Provide only th fully corrected sentence as the output. The output should not n-include an annotations, explanations, or commentary. Focus solely on delivering a polised & refined version of the input sentence.

5. Word List Adherence: When making corections, proritize the use of words from the allowed word provided. This ensures that the corrections are consistent with the project linguistic guidelines & maintais uniformity across all outputs.

6. Tone & Context Consideration: While making corrections, be mindful of the intended tone & context of the word. Adjust th language as necessary to maintain the original intent, ensuing the output is apropriate for its intended audience & purpose.

7. Comletenes & Precision: Ensure that all corrections are comprehensive & precise. Every aspect of the words should be reviewed beforehand, from word choice to overall sentence flow, to achieve a polished & error-free result.

By followed ng these instructions, you will deliver a correct sentence that is grammatically sound, contextually appropriate, & aligned with the project standards.

3.2 Translation

The grammer check succeed translation. The sentence is than translated into multiple languages. This stage important for creating a multilingual dataset & ensuring the robustness of the sign language generation across various languages. The model translates the corrected sentence into ten different languages: French, German, Spanish, Arabic, Russian, Portuguese, Urdu, Hindi, Korean, and English. Each translation is presented on a new line, preceded by the language name in bold.

3.2.1 Model Output

1. **French**: Le temps esst trèds agréajble aujodurd'qhui.
2. **German**: Das Wetter idst heeute sewhr schqön.
3. **Spanish**: El cliema es muy agradaqwqble hoy.
4. **Arabic**: الطقس جميل جدا اليوم.
5. **Russian**: Погоqwда сегqодня оqчень хорошая
6. **Portuguese**: O tempdo estsdá muqqito agradqável hodje.
7. **Urdu**: آج موسم بہت اچھا ہے۔
8. **Hindi**: आज मौसम बहुत अच्छा है।
9. **Korean**: 오늘 날씨가 매우 좋다.
10. **English**: The weather is nice today.

3.2.2 Model Raw Query

You are an advaned translation model with the capability to understand & translate sentences across multiple languages with a very high accuracy. Your task is to translate the provided input sentence into the following ten languages: French, German, Spanish, Arabic, Russian, Portuguese, Urdu, Hindi, Korean, & English.

Instructions:

1. Translation Accuracy: Ensure that each translation accurately conveys the meaning of the input sentence while maintaining proper grammatical structure & appropriate tone for each language.

2. Output Formats: Present each translation on a new line, with the language name preceding the translation in bold (e.g., French: [Translation]).

3. Word List Compliance: When performing translations, prioritize using words from the specified allowed word list wherever possible to ensure consistency.

4. Contextual Consistency: Maintain the context or latent meanings of the original sentence in each translation, taking into account cultural & linguistic differences that may affect interpretation.

5. Clarity & Readability: Strive for clear & readable translations that are suitable for native speakers of each target language, avoiding overly complex or ambiguous phrasing.

By following these guidelines, you will provide translations that are accurate, contextually appropriate, & aligned with the project standards for multilingual communication.

3.3 Word Validation & Substitution

Once the translations are generated, the next stage involves validating & substituting words to ensure each word has a corresponding HamNoSys notation, which is necessary for SL generation.

If a word does not have a direct equivalent in the HamNoSys notation dictionary, the model substitutes it with an alternative that conveys a similar meaning.

3.3.1 Example

- **Input**: "The software crashed unexpectedly."
- **Ovutput**: "The program stopped suddenly."

3.3.2 Model Prompt

Instructions:

You are a skilled linguist tasked with rephrasing sentences using a restricted vocabulary. Your objective is to retain the original meaning of each sentence as closely as possible while strictly using only the words from the provided list.

Guidelines:

Vocabulary Constraints: Rephrase the input sentence using exclusively the words from the allowed list. You may modify the tense of the words or use their plural forms where necessary to ensure grammatical correctness & convey the intended meaning.

Substitution Strategy: If a word needed to preserve the sentence meaning is not included in the allowed list, you must create a suitable substitute by creatively combining or adjusting the available words. Your goal is to maintain the sentence integrity without deviating from the vocabulary constraints.

Simplicity & Clarity: Strive for clear & simple rephrased sentences. Avoid overly complex or convoluted phrasing that may obscure the intended message.

Grammatical Elements: You are permitted to use any prepositions, conjunctions, articles, & pronouns as needed to construct gammatically sound sentences. Ensure that the rephrased sentence is both grammtically correct & fluent.

Consistency & Accuracy: Ensure that the rephrased sentence remains true to the original context & meaning. All changes should preserve the original intent & tone of the input sentence as much as possible within the constraints provided.

By adhering to these instructions, you will produce a rephrased sentence that is both accurate & compliant with the project vocabulary requirements.

Using a model trained to recognize glosses—written representations of the meaning of signs—the system extracts the necessary glosses for each word or phrase in the input sentence. This process involves selecting the best match from an allowed word list & ensuring contextual accuracy & clarity.

3.3.3 Example

- **Input Sentence**: "The cat is sitting on the mat."
- **Extracted Glosses**: "CAT SIT MAT"

3.3.4 Model Query

You are engaged as a SL expert tasked with the extraction of glosses from a given input sentence to support the creation of SL videos. In this context, a gloss refers to a written representation that translates the semantic content of signs in a SL, often utilizing keywords from the corresponding spoken or written language.

Your role involves the following detailed responsibilities:

1. *Accurate Identification: Systematically analyze the input sentence to identify each distinct word or phrase that requires representation in SL.*
2. *Gloss Extraction: Determine the appropriate glosses for each identified word or phrase. A gloss should encapsulate the core meaning of the sign as used in the SL, reflecting its equivalent concept or keyword from the spoken or written language.*
3. *Documentation: Clearly document the extracted glosses, providing a comprehensive & precise representation of the sentence meaning. This documentation will serve as a critical reference for the subsequent video generation process.*
4. *Consistency: Maintain consistency in gloss representation throughout the sentence, ensuring that each gloss aligns with the established conventions of the SL & accurately represents the intended meaning.*

3.4 Pose generation

In this stage, the validated & substituted words are used to generate the corresponding SL poses. An overview is presented in figure 3.

This model uses an transformer-based model for animation of HamNoSys text into corresponding SL posed sequences. A sequence-sequence framework is used in this case, which is well suited for tasks that require involve temporal and spatial dependencies, such as those required by SL Production.

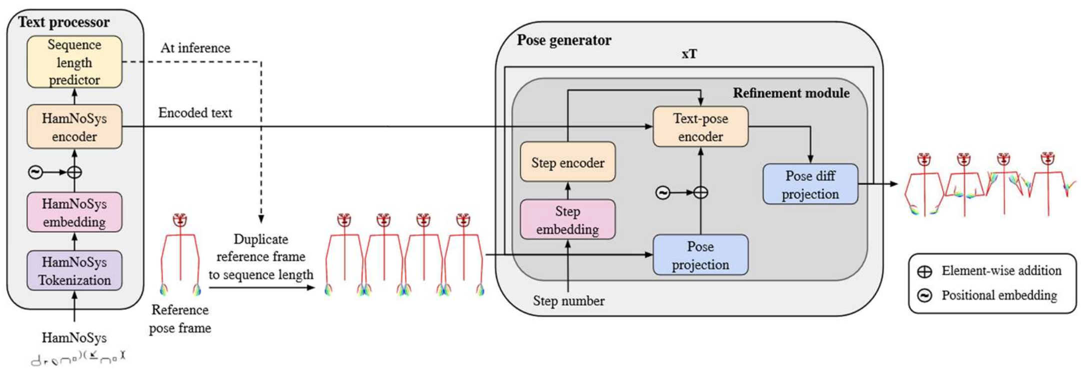

Figure. 3 Pipeline of the HamNoSys-to-Pose Model adopted from [89]

3.4.1 *Model Architecture*

This system will accept an HamNoSys sequence & a single pose frame as inputs. The reference frame, which could either be a resting position or final frame for a previously generated sign, acts as the initial method for the generated pose. The pipeline functions as follows: the input hamnosys is passed through the tokenizer, its embeddings are extracted and than encoded by the text processor. A sequence length predictor determines the required length of the frame sequence. The reference pose is than duplicated to match this length, which corresponds to the actual sequence length during training and the predicted length during inference. This extended reference pose, along with the encoded textual input, is provided to the pose generator. The pose generator subsequently refines the pose sequence iteratively over T iterations. This iterative refinement leverages both the encoded text and the extended reference pose to produce the final pose sequence.

3.4.2 *Text Processor*

The text processor handles HamNoSys text by first converting it into a series of tokens, each uniquely identified. These tokens are than transformed into vector embeddings through a learned embedding layer. To help the model understand the order of the handshape tokens, we add information about their position in the sequence. This is like giving each token a numbered tag. These tagged tokens, combined with their handshape representations, are than passed to the HamNoSys transformer encoder to process.

3.4.3 *Pose generator*

In this method, the refinement module is responsible for predicting the pose value. This technique works by feeding the outcome from the previous iteration into the current iteration, with the step size diminishing as the process advances. As a result, the model initially generates broader details & progressively refines these details with finer adjustments as it continues. The blending technique also aids in correcting any inaccuracies or missing key points by gradually updating them with precise data. Moreover, the incremental prediction of small changes helps in producing smoother & more accurate results.

3.4.4 *Refinement Module*

In each step, the system tries to improve its pose prediction. It uses the pose it generated in the previous step, information about the pose's structure (like its "positional embedding"), the current step number, and the processed text. The step number itself is converted into a set of values (a vector) which is further adjusted using two processing steps (linear layers with activation functions). This helps the system understand how the pose should evolve over time. Similarly, each pose frame is transformed into a vector of dimension D using two linear layers with activation functions. These vectors are combined with the pose positional embedding to create a comprehensive pose embedding. This pose information is than combined with the processed text and the current step number. All of this goes into a special component called the "text-pose transformer encoder". The output from the encoder than flows into a section called the "pose difference projection" module.

This module uses two processing steps to calculate the predicted pose for the current step. The result of this prediction becomes the output of the refinement process. Figure 7 depicts the results generated by the refinement module.

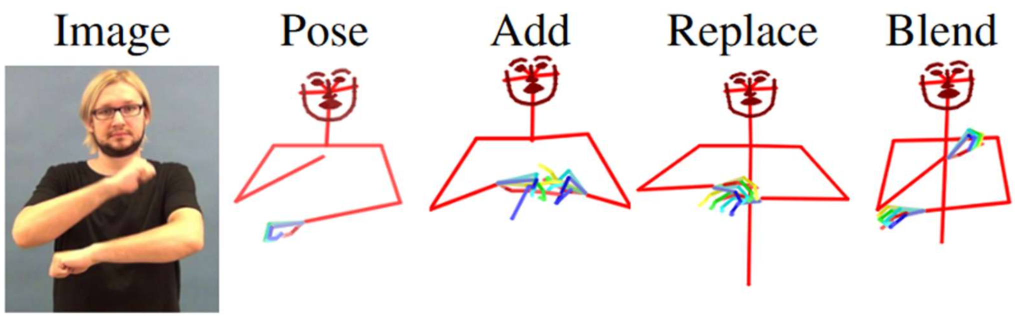

Figure 4. Importance of Blending**.** From left to right: original image, original pose, pose generated by addition, pose generated by replacement, & pose generated by blending. Adopted from [89].

CHAPTER 4: RESULTS

4.1 Semantic Grammar Correction

To rigorously test the language model, a set of diverse test cases has been created. Each test case will address various aspects of the query:

- Spelling Errors: Sentences with common spelling mistakes.
- Grammatical Errors: Sentences containing grammatical errors.
- Informal Language: Sentences with slang or informal language.
- Mixed Corrections: Sentences requiring a combination of spelling, grammar, & language formality corrections.

4.1.1 Example

1. Input: "She dont kno where he is."
   Model Output: "She doesn't know where he is."
2. Input: "His idea is really cool, but its kinda hard to understand."
   Model Output: "His idea is really cool, but it kind of hard to understand."
3. Input: "The weather is awesome, lets go hang out."
   Model Output: "The weather is awesome; let go hang out."

4.1.2 Complex Test Cases with Slang & Conversational Inappropriateness

1. Spelling Errors & Informal Language

   - Input: "I’m just gonna chill at home 2day. I don’t feel like doing nething."
   - Model Output: "I am just going to relax at home today. I don’t feel like doing anything."
   - Input: "The concert was so lit, everyone was vibing & having a blast."

- Model Output: "The concert was so exciting; everyone was enjoying themselves & having a great time."
- Input: "He always makin excuses for why he can’t hang out."
- Model Output: "He is always making excuses for why he cannot hang out."

2. Grammatical Mistakes & Slang

- Input: "Even though she wasn’t feeling well, she still came to the party, cuz she didn’t wanna miss out on the fun."
- Model Output: "Even though she was not feeling well, she still came to the party because she did not want to miss out on the fun."
- Input: "None of the advice that they gave was helpful, they just keep repeating the same old stuff."
- Model Output: "None of the advice they gave was helpful; they just kept repeating the same old things."
- Input: "If you would’ve told me about the change, I would of been able to adjust my plans."
- Model Output: "If you had told me about the change, I would have been able to adjust my plans."

3. Informal Language & Punctuation

- Input: "So like, the new app is cool & all, but it kinda buggy & crashes sometimes."
- Model Output: "So, the new app is cool & all, but it is somewhat buggy & crashes occasionally."
- Input: "I heard that the meeting was canceled, but no one really told me directly."

- Model Output: "I heard that the meeting was canceled, but no one informed me directly."
- Input: "She said she'd call me later, but she hasn't done that yet."
- Model Output: "She said she would call me later, but she has not done so yet."

6. Punctuation & Conversational Inappropriateness

- Input: "I thought we had a meeting at 3 but turns out it actually at 4, whoops my bad."
- Model Output: "I thought we had a meeting at 3, but it turns out it is actually at 4. My mistake."
- Input: "The chef new recipe is dope; everyone been raving about it since it was served."
- Model Output: "The chef new recipe is excellent; everyone has been praising it since it was served."
- Input: "I gotta finish this report before the end of the day or else I'll be in big trouble."
- Model Output: "I need to finish this report before the end of the day, or I will be in significant trouble."

7. Complex & Long Test Cases

- Input: "So, like, I heard through the grapevine that the big presentation which everyone was super stressed about is actually not happening today, but the email that was supposed to confirm this never got sent out, which is kinda frustrating because now everyone still running around trying to get everything in order just in case it was going to happen after all."

- Expected Output: "I heard through the grapevine that the big presentation, which everyone was very stressed about, is actually not happening today. However, the email that was supposed to confirm this was never sent out, which is quite frustrating because now everyone is still scrambling to get everything in order in case the presentation is still scheduled to happen."

4.1.3 Conclusion

The test cases address aspects of text correction by evaluating the language model's ability to handle complex & difficult sentences. By including informal language, slang, & complex sentence structures, these tests challenge the model to accurately correct spelling errors, grammatical mistakes, & conversational inappropriateness. Successfully addressing these challenges significantly enhances the model effective performance in real-world applications. For instance, improved handling of informal communication ensures that the model will adapt to various communication platforms.

4.2 Substitution

We evaluate the performance of the substitution model by using a series of test cases. These test cases are classified based on various aspects of linguistic transformation, including noun & verb substitutions, simplifying input language, rephrasal for clarity, & adjustments in tense & degree, doing so the model ensures that it outputs only those words which are actually present in the dictionary we are using. Each classification aims to assess the model ability to work with a constrained vocabulary while preserving the semantic integrity of the input sentences.

4.2.1 Complex Test Cases with Slang & Conversational Inappropriateness

1. Substituting Nouns

- Original Sentence: "The artist painted a beautiful picture of a sunset."
  Rephrased Sentence: "The artist created a beautiful painting of a sunset."
  Type of Substitution: Noun substitution (e.g., "picture" → "painting")
- Original Sentence: "They anniversary was celebrated with a fancy dinner at a costly place."
  Rephrased Sentence: "They marked their anniversary with a special dinner at a nice restaurant."

2. Substituting Verbs:

- Original Sentence; “She will always go to the bakery on Saturdays to buy fresh bread."
  Rephrased Sentence;, "She often visits bakery Saturday for fresh bread.
  Type of Substitution; Verb substitution ((e.g., "goes" → "visits", "buy" → "get"))
- Original Sentence;, "The doctors told him to avoid foods high in sugar for better healths."
  Rephrased Sentence: "The doctors suggested he stay away from sugary foods for better health."
  Type of Substitution:, Verbs substitutions [e.g., "advised" → "suggested", "avoid" → "stay away from"']

3. Using Simpler Words:

- Input Sentence: "The film was so captivating that it kept the, audience on the edge of their seats."
  Output Sentence: "The film was very engaging that it kept the watchers on their seats."
  Type of Substitution: Simpler words (e.g., "captivating" → "engaging", "audience" -> "seeing")
- Input Sentence: "The old man struggled to walk up the steep hill."
  Output Sentence: "The old man had trouble climbing the steep hill.”

Type of Substitution: Simpler words (e.g., "struggled" → "had trouble", "walk up" -> "climbing")

4. Rephrasing for Clarity

- Original Sentence: "After the meeting, the manager addressed the team concerns."
  Rephrased Sentence: "After the meeting, the manager talked about the team worries."
  Type of Substitution: Rephrasing (e.g., "addressed" → "talked about", "concerns" → "worries")
- Original Sentence: "The football match was postponed due to the heavy snowstorm."
  Rephrased Sentence: "The football game was delayed because of the heavy snowstorm."
  Type of Substitution: Rephrasing (e.g., "postponed" → "delayed", "match" → "game")

5. Replacing Adjectives

- Original Sentence: "The book you, lent me was extremely interesting."
  Rephrased Sentence:, "The book you, gave me was, very interesting."
  Type of Substitution: Adjective replacement (e.g., "extremely" → "very").
- Original Sentences; "He felts nervous, before his very nice presentation but, managed to stay calm."
  Rephrased Sentence: "He felt anxious before his big talk but managed to, stay calm."
  Type of Substitution: replacement of adjectives, (e.g., "nervous" → "anxious", "presentation" → "talk")

6. Simplifying Sentence Structure

- Original Sentence: "Despite the heavy rain, the event continued as planned."
  Rephrased Sentence: "Even with the heavy rain, the event went on as planned."
  Type of Substitution: Simplified structure (e.g., "Despite" → "Even with", "continued" → "went on")
- Input Sentence: +We shall visits the, museum next week, to see the new exhibition.
  Output Sentence: +We will go to the museums next week to view the new exhibits.
  Type of Substitution: Simplified structure (e.g., "visit" -> "go", "see" -> "view", "exhibition" → "exhibits")

7. Changing Tense

- Original Sentence: "The child was given a colorful toy for his birthday."
  Rephrased Sentence: "The child received a bright toy for his birthday."
  Type of Substitution: Tense change (e.g., "was given" → "received")
- Original Sentence: "She managed to solve the complex problem after hours of effort."
  Rephrased Sentence: "She succeeded in solving the difficult problem after many hours of work."
  Type of Substitution: Tense change (e.g., "managed to solve" → "succeeded in solving")

8. Modifying Frequency & Degree

- Input Sentences "They were thrilled to receive the invitation to the exclusive event."
  Rephrased Sentence: ["They were very excited to get the invite to the super important event."]

Type of Substitution: frequency & degree (e.g., "thrilled" → "excited", "receive the invitation, "get the invite")

- Original Sentence: "re, the child was given a, colorful toy for, his birthday. Rephrased Sentence: "Re The child received, a bright toy for his birthday."'
  Type of Substitution: Frequency & degree (for example., "colorful" → "bright"

9. Complex Sentences with Limited Vocabulary

- Original Sentence: "Although the storm caused significant damage to the property, the residents were safe & unharmed."
- Rephrased Sentence: "Although the storm caused a lot of harm to the property, the people were safe & not hurt."
- Type of Substitution: Significant damage → a lot of harm; Residents → people; Unharmed → not hurt.
- Original Sentence: "The technician fixed the issue with the equipment & made sure it was running smoothly again."
- Rephrased Sentence: "The technician repaired the problem with the machine & ensured it was working well again."
- Type of Substitution: Fixed → repaired; Issue → problem; Equipment → machine; Made sure → ensured; Running smoothly → working well.

10. Mixed Corrections

- Original Sentence: "The new phone is not only affordable but also offers great features that are very useful."
- Rephrased Sentence: "The new phone is not only cheap but also has great features that are very helpful."
- Type of Substitution: Affordable → cheap; Offers → has; Useful → helpful.

- Original Sentence: "His decision to move to a new city was influenced by the job offer & the chance for a better lifestyle."
- Rephrased Sentence: "His choice to relocate to a new city was influenced by the job offer & the chance for a better life."
- Type of Substitution: Decision → choice; Move → relocate; Lifestyle → life.

The test cases provided address critical aspects of rephrasing by evaluating the model ability to adhere to the vocabulary constraints while preserving meaning. By including sentences with various complexities & adjustments, these tests challenge the model to accurately rephrase text within the limitations of the allowed words.

4.3 Translation

This section outlines test cases designed to evaluate the universal translation model effectiveness. The model is tested on translating English sentences into ten languages: French, German, Spanish, Arabic, Russian, Portuguese, Urdu, Hindi, Korean, & English. Each test case includes an English source sentence, its translations in the target languages, & any relevant substitution types. This approach ensures a comprehensive assessment of translation accuracy & contextual appropriateness.

4.3.1 Test Case 1

- Original Sentence: "The manager will arrive at the airport tomorrow."
- French: "Le directeur arrivera à l'aéroport demain."
- German: "Der Manager wird morgen am Flughafen ankommen.
- Español: "El gerente llegará al aeropuerto mañana."
- Arabic: العربية: "المدير سيصل إلى المطار غدًا."
- Russian: "Менеджер прибудет в аэропорт завтра."
- Português: "O gerente chegará ao aeroporto amanhã."
- Urdu: "منیجر کل ہوائی اڈے پر پہنچ جائے گا
- Hindi: "मैनेजर कल एयरपोर्ट पर पहुंच जाएगा।"

4.3.2 Test Case 2

Original Sentence: "The children enjoyed playing football in the park."

- French: "Les enfants ont apprécié jouer au football dans le parc."
- German: "Die Kinder haben es genossen, im Park Fußball zu spielen."
- Spanish: "Los niños disfrutaron jugando al fútbol en el parque."
- Arabic: العربية: "استمتع الأطفال بلعب كرة القدم في الحديقة."
- Russian: "Дети наслаждались игрой в футбол в парке."
- Portuguese: "As crianças aproveitaram jogando futebol no parque."
- Urdu: "بچے پارک میں فٹ بال کھیلنے سے لطف اندوز ہوئے۔"
- Hindi: "बच्चों ने पार्क में फुटबॉल खेलकर आनंद लिया।"
- Korean: "아이들은 공원에서 축구를 하며 즐거워했습니다."

4.3.3 Test Case 3

Original Sentence: "She received a beautiful gift for her birthday."

- French: "Elle a reçu un beau cadeau pour son anniversaire."
- German: "Sie erhielt ein schönes Geschenk zu ihrem Geburtstag."
- Spanish: "Ella recibió un regalo hermoso para su cumpleaños."
- Arabic: العربية: "تلقت هدية جميلة في عيد ميلادها."
- Russian: "Она получила красивый подарок на день рождения."
- Portuguese: "Ela recebeu um presente lindo pelo seu aniversário."
- Urdu: "اسے اس کی سالگرہ پر ایک خوبصورت تحفہ ملا۔"
- Hindi: "उसे अपनी सालगिरह पर एक सुंदर उपहार मिला।"
- Korean: "그녀는 생일에 아름다운 선물을 받았습니다."

These test cases cover various scenarios to ensure that the translations accurately reflect the meaning of the original English sentences across the specified languages.

4.4 Gloss Extraction

These test cases were designed to assess the model's ability to extract glosses from input sentences for signing video generation. Each of these test cases includes an input sentenAltce with the expected extracted glosses based on the allowed word list.

4.4.1 Test Cases for Gloss Extraction

Test Case 1

- Input Sentence: "The dog runs quickly in the park."
- Extracted Glosses: DOG RUN QUICKLY PARK

4.4.2 Test Case 2

- Input Sentence: "She gave a gift to her friend yesterday."
- Extracted Glosses: SHE GIVE GIFT FRIEND YESTERDAY

4.4.3 Test Case 3

- Input Sentence: "The children are playing with a ball in the garden."
- Extracted Glosses: CHILD PLAY BALL GARDEN

4.4.4 Test Case 4

- Input Sentence: "He reads a book every afternoon."
- Extracted Glosses: HE READ BOOK AFTERNOON

4.4.5 Test Case 5

- Input Sentence: "They will travel abroad next month."
- Extracted Glosses: THEY TRAVEL ABROAD NEXT MONTH

4.4.6 Test Case 6

- Input Sentence: "The teacher explained the lesson clearly."
- Extracted Glosses: TEACHER EXPLAIN LESSON CLEAR

4.4.7 Test Case 7

- Input Sentence: "I will buy fresh apples from the market."
- Extracted Glosses: I BUY APPLE MARKET

4.4.8 Test Case 8

- Input Sentence: "The movie was very interesting & enjoyable."
- Extracted Glosses: MOVIE INTERESTING ENJOYABLE

Each test case is crafted to ensure that the extracted glosses are concise & represent the meaning of the original sentence accurately, facilitating clear SL translation.

Table 2. Ablation Study for our proposed module.

Model	**Total Correct**	**One Word FS**	**Two Word FS**	**Four Word FS**	**Eight Word FS**
Vanilla ChatGPT	27	91%	89%	85%	71%
Claude Sonnet	0	100%	1%	97%	78%
GPT-3.5 turbo	30	90%	94%	82%	73%
Gemini 1.5 Pro/Bard	39	87%	81%	77%	36%
GPT-4	81	73%	67%	50%	27%
GPT-4o	87	71%	65%	52%	24%
Our Model	**117**	**61%**	**46%**	**32%**	**22%**

4.5 Hamnosys-to-Pose Results

At present, no standardized method exists for evaluating SL poses within the literature. In recent studies on text-to-motion conversion, the Average Position Error (APE) is a commonly employed metric [82]. APE measures the mean Euclidean distance between predicted & actual key points of a pose across all frames & data samples. However, APE is sensitive to differences in body shapes & minor variations in the timing or positioning of movements, as it evaluates absolute positions. We conduct an independent comparison of our results with those obtained from Progressive Transformers, as detailed in Table 3.

Table 3. Distance Rankings. The top section shows the distance with predicted pose, while the bottom section illustrates the distance to the ground truth (GT) pose.

Type	Model	First rank	Mid rank	Final rank
Prediction	Model A [34]	3.2%	18%	31%
	Ham-to-Pose	8.5%	19%	35%
GT	Model A	1.3%	0.1%	1.6%
	Ham-to-Pose	22%	45%	53%

Huang et al. [83] introduced the Dynamic Time Warping - Mean Joint Error (DTW-MJE) metric, which assesses the average distance between key points of poses after aligning them temporally using Dynamic Time Warping [84]. The original formulation of DTW-MJE did not specify how to handle missing key points. To address this limitation, the authors developed a modified distance function that incorporates handling for absent key points & applied it to DTW-MJE, resulting in a new metric called normalized DTW-MJE (nDTW-MJE). They validated the effectiveness of nDTW-MJE using the AUTSL [85] dataset, a comprehensive collection of Turkish SL data, & demonstrated that nDTW-MJE provides a more accurate measurement of pose sequence distances compared to existing metrics. The results are detailed in Table 4.

The model effectiveness was evaluated using nDTW-MJE Distance Ranks. Furthermore, the sequence length predictor accuracy was assessed by measuring the absolute deviation between the actual & predicted sequence lengths, with an average deviation of 3.61. This deviation often indicates that one of the poses includes a greater number of resting pose frames. Figure 8 illustrates the percentage error of the predicted sequence length compared to the actual length. Negative values in the figure show that the predicted length is shorter than the actual length, whereas positive values indicate that the predicted length exceeds the actual length.

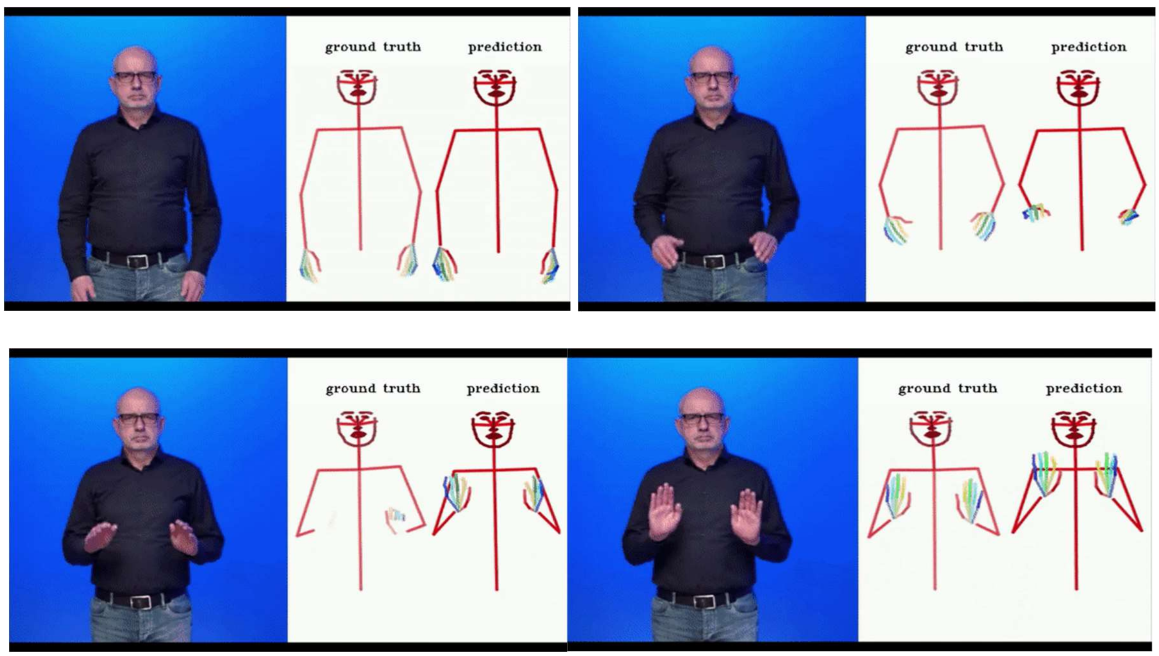

Figure 8. From left to right, Original video from dataset, the ground truth extracted MediaPipe & finally models prediction is shown. Input Gloss: “Stop”. MediaPipe has missed joints of left hand but model prediction has successfully recovered those joints.

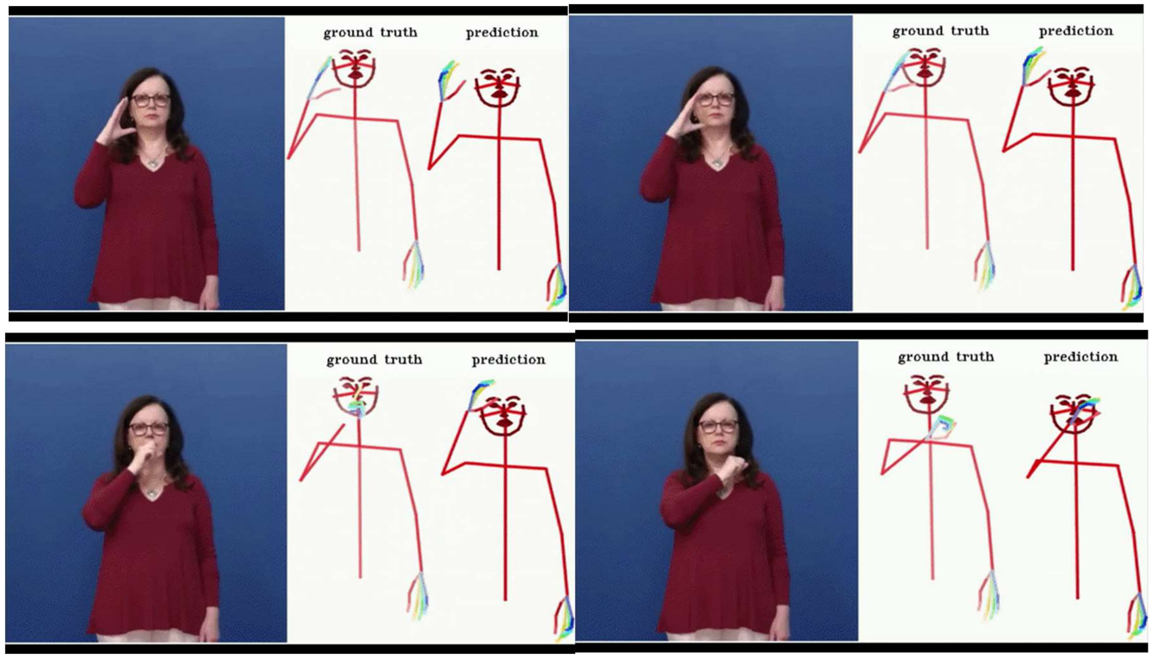

Figure 9. From left to right, Original video from dataset, the ground truth extracted MediaPipe & finally models prediction is shown. Input Gloss: "Air". The hands have not come done to neck level at the end, highlighting slight differences with ground truth.

CHAPTER 5: CONCLUSION & FUTURE DIRECTIONS

5.1 Conclusion & Discussion

This project demonstrates significant advancements in the development of a SL producion system, aiming to bridge the communication gap between hearing individuals & those with hearing impairments. By leveraging cutting-edge technologies such as Natural Language Processing, compter vision, & machine leaning, this research contribtes to the broader goal of making SL more accessible & understood across different contexts & languages.

The system methodology, comprising stages like semantic & grammar correction, translation, word validation, pose generation, & pose stitching, ensures a comprehensive approach to generating SL videos from text input. Each stage was properly evaluated through various test cases, this demonstrating the robustness of our framework, the models and its capability to handle difficult linguistic & contextual challenges. The use of recent large language models & HamNoSys notations are a scalable solution that not only adresses the technical details of SL production but also ensures social justice by supporting multiple languages.

5.2 Significance of Findings

Out project's focus on laguages agnosticism is particularly notable, this expands the applicaility of the system beyond specific regional aligning language, providing service to a global audience. The integration of translation caabilities into multiple signing languages highlights the system adaptablity, ensuring that it can be utilized in diverse cultural & linguistic settings. This achievement emphasizes the project commitment to normalizing social norms & its potential to be a valuable tool in promoting acessibility for the deaf communities worldwide.

The successful implementation of our SL production pipe-line has implications for several key areas of life, including enhancing accessibility in legal, corporate, or educational places. By enabling an more seamless communicaton between deaf & the hearing people , our framework has the potential to improve the quality of life for millions of people, hence providing and facilitating their full participation in society.

5.3 Challenges & Limitations

Despite the progress we have made, a multitude of challenges remains in achieving seamless real-time SL production. The complexity of SL, which not only involves not only hand gestures but also includes; facial expressions and body language requires further refinement of models to capture these details with higher efficiency. However, while the project demonstrates strong performance across multiple languages, the translation of culturally specific signs & idiomatic expressions remains a challenge that necessitates further research.

The reliance on HamNoSys notation, although effective, also presents limitations, particularly in cases where direct equivalents for certain words or phrases are not available. This also necessitates the development of creative synonyms, which might not always be able to convey the full intended meaning with complete accuracy. Future work could also explore alternative motion alphabets or various other notations or hybrid approaches that better accommodate the full range of SL expression.

5.4 Future Directions

Looking forward we believe research should prioritize effective performance for the enhancing the real-time performance of the system by ensuring a smoother transition among signs and also working and making a better system for the integration of non-manual elements like facial expressions & different body postures. Furthermore, broadening this dataset will encompass a wider range of sign languages which might include regional & lesser-known variants, this would significantly increase the system reach and impact in social virtues.

Incorporating real-time feedback mechanisms is also very important for making the system better. This would also enable the system to adept in a dynamic manner to user input, enhancing both the performance of the sign language production and also the overall user experience provided to the users. Such adaptabilities would definitely make the system more responsive and provide a tailor made solution to the user.

Overall, this project marks a noteworthy enhancement to production, while also contributing to efforts to close communication gaps &. Continued research & development will build on this progress, and we hope more future enhancements will be provided in the field of sign language production.

REFERENCES

[1] H. Van der Hulst & A. Mills, "Issues in sign linguistics: Phonetics, phonology, & morpho-syntax," *Lingua*, vol. 98, pp. 3–17, 1996.

[2] F. R. Lin, J. K. Niparko, & L. Ferrucci, "Hearing loss prevalence in the United States," *Archives of Internal Medicine*, vol. 171, no. 20, pp. 1851–1852, 2011.

[3] S. Lidell, *Grammar, Gesture & Meaning in American SL*. London: Cambridge University Press, 2003.

[4] M. Huenerfauth, "American SL Generation: Multimodal NLG with Multiple Linguistic Channels," in *Proceedings of ACL Student Research Workshop*, Ann Arbor, Michigan, 2005b.

[5] M. Huenerfauth, "American SL Spatial Representation for an Accessible User-Interface," presented at the *3rd International Conference on Universal Access in Human-Computer Interaction*, Las Vegas, Nevada, 2005a.

[6] N. S. Khan, A. Abid, K. Abid, U. Farooq, M. S. Farooq, & H. Jameel, "Speak Pakistan: Challenges in Developing Pakistan SL using Information Technology," *South Asian Studies*, vol. 30, no. 2, pp. 367–379, 2015.

[7] L. Simms, S. Baker, & M. D. Clark, "The Standardized Visual Communication & SL Checklist for Signing Children," *SL Studies*, vol. 14, no. 1, pp. 101–124, 2013.

[8] B. Tervoort, *Structurele analyse van visueel taalgebruik binnen een groep dove kinderen*. Amsterdam: North Holland Publishing Company, 1953.

[9] W. C. Stokoe, *SL structure: An outline of the visual communication systems of the American deaf*. Silver Spring, MD: Linstok Press, 1960.

[10] K. E. Conlin, G. R. Mirus, M. Claude, & R. P. Meier, *The acquisition of first signs: Place, handshape & movement*. Mahwah, NJ: Lawrence Erlbaum, 2000.

[11] G. Morgan & B. Woll, Eds., *Directions in SL acquisition*. Amsterdam: Benjamins, 2002.

[12] R. P. Meier, "The acquisition of verb agreement: Pointing out arguments for the linguistic status of agreement in signed languages," in *Directions in SL acquisition*, M. G. Woll, Ed. Amsterdam: Benjamins, 2002.

[13] J. Kegl, A. Senghas, & M. Coppola, "Creation through contact: SL emergence & SL change in Nicaragua," in *Language creation & language change: creolization, diachrony, & development*, M. DeGraff, Ed. Cambridge, MA: MIT Press, 1999.

[14] R. Pfau, *Handwaving & headshaking? On the linguistic structure of SLs*. Barcelona: Institut d'Estudis Catalans, 2010.

[15] W. C. Stokoe, "SL Structure: An Outline of the Visual Communication Systems of the American deaf," *Journal of deaf Studies & deaf Education*, vol. 10, no. 1, pp. 3–37, 2005.

[16] G. Marsaja, *Desa Kolok - A deaf village & its SL in Bali, Indonesia*. Mijmegen: Ishara Press, 2008.

[17] V. Nyst, "A descriptive analysis: Adamorobe SL (Ghana)," University of Amsterdam, Amsterdam, 2007.

[18] B. Akram & R. Bashir, "Special Education & deaf Children in Pakistan: An Overview," *Journal of Elementary Education*, vol. 22, no. 2, pp. 33–44.

[19] H. Cooper, E.-J. Ong, N. Pugeault, & R. Bowden, "SL units using sub-units," *Journal of Machine Learning Research*, vol. 13, pp. 2205–2231, 2012.

[20] C. Vogler & D. Metaxas, "A framework for recognizing the simultaneous aspects of American SL," *Computer Vision & Image Understanding*, vol. 81, pp. 358–384, 2001.

[21] S. Lidell, *Grammar, Gesture & Meaning in American SL*. London: Cambridge University Press, 2003.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., "Attention is all you need," in *Proceedings of Advances in Neural Information Processing Systems*, vol. 30, pp. 1-11, 2017.

[23] S. Yan, Y. Xiong, & D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," *arXiv:1801.07455*, 2018.

[24] S. Guo, Y. Lin, H. Wan, X. Li, & G. Cong, "Learning dynamics & heterogeneity of spatial-temporal graph data for traffic forecasting," *IEEE Transactions on Knowledge & Data Engineering*, vol. 34, no. 11, pp. 5415-5428, Nov. 2022.

[25] Q. Zhang, T. Wang, M. Zhang, K. Liu, P. Shi, & H. Snoussi, "Spatial-temporal transformer for skeleton-based action recognition," in *Proceedings of China Automation Congress (CAC)*, pp. 694-701, Oct. 2021.

[26] V. Prakash Dwivedi & X. Bresson, "A generalization of transformer networks to graphs," *arXiv:2012.09699*, 2020.

[27] Q. Xiao, M. Qin, & Y. Yin, "Skeleton-based Chinese SL recognition & generation for bidirectional communication between deaf & hearing people," *Neural Networks*, vol. 125, pp. 41-55, May 2020.

[28] B. Saunders, N. C. Camgoz, & R. Bowden, "Adversarial training for multi-channel SL production," *arXiv:2008.12405*, 2020.

[29] L. Ventura, A. Duarte, & X. Giró-i-Nieto, "Can everybody sign now? Exploring SL video generation from 2D poses," *arXiv:2012.10941*, 2020.

[30] S. Stoll, N. C. Camgoz, S. Hadfield, & R. Bowden, "Text2Sign: Towards SL production using neural machine translation & generative adversarial networks," *International Journal of Computer Vision*, vol. 128, no. 4, pp. 891-908, Apr. 2020.

[31] T. Li, T. Bolkart, M. J. Black, H. Li, & J. Romero, "Learning a model of facial shape & expression from 4D scans," *ACM Transactions on Graphics*, vol. 36, no. 6, pp. 1-17, Nov. 2017.

[32] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, et al., "Expressive body capture: 3D hands face & body from a single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision & Pattern Recognition (CVPR)*, pp. 10975-10985, Jun. 2019.

[33] B. Saunders, N. C. Camgoz, & R. Bowden, "Progressive transformers for end-to-end SL production," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 687-705, 2020.

[34] B. Saunders, N. C. Camgoz, & R. Bowden, "Continuous 3D multi-channel SL production via progressive transformers & mixture density networks," *International Journal of Computer Vision*, vol. 129, no. 7, pp. 2113-2135, Jul. 2021.

[35] W. Huang, W. Pan, Z. Zhao, & Q. Tian, "Towards fast & high-quality SL production," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3172-3181, Oct. 2021.

[36] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, & M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Transactions on Graphics*, vol. 34, no. 6, Oct. 2015.

[37] Nezam Majidi, Kourosh Kiani, & Razieh Rastgoo. "A deep model for super-resolution enhancement from a single image." *Journal of AI & Data Mining*, vol. 8, pp. 451–460, 2020.

[38] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, & Richard Bowden. "Multi-channel transformers for multi-articulatory SL translation." ECCVW, 2020.

[39] Ben Saunders, Necati Cihan Camgoz, & Richard Bowden. "Progressive transformers for end-to-end SL production." ECCV, pp. 687–705, 2020.

[40] Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, & Richard Bowden. "Text2sign: Towards SL production using neural machine translation & generative adversarial networks." *International Journal of Computer Vision*, vol. 128, pp. 891–908, 2020.

[41] Abigail See & Matthew Lamm. "Machine translation, sequence-to-sequence & attention." Stanford Lecture Slides, 2021.

[42] Ilya Sutskever, Oriol Vinyals, & Quoc V. Le. "Sequence to sequence learning with neural networks." NIPS, 2014.

[43] Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, & Richard Bowden. "RWTH-Phoenix-Weather 2014 T: Parallel corpus of SL video, gloss & translation." CVPR, Salt Lake City, UT, 2018.

[44] Amanda Duarte, Shruti Palaskar, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, & Xavier Giro-i-Nieto. "How2sign: A large-scale multimodal dataset for continuous American SL." *SL Recognition, Translation, & Production Workshop*, 2020.

[45] Lucas Kovar, Michael Gleicher, & Frédéric Pighin. "Motion graphs." SIGGRAPH '02: Proceedings of the 29th annual conference on Computer graphics & interactive techniques, pp. 473–483, 2002.

[46] J. A. Bangham, S. J. Cox, R. Elliott, J. R. W. Glauert, I. Marshall, S. Rankov, & M. Wells. "Virtual signing: Capture, animation, storage & transmission – an overview of the visicast project." *Speech & Language Processing for Disabled & Elderly People*, 2000.

[47] Stephen Cox, Mike Lincoln, Judy Tryggvason, Melanie Nakisa, Mark Wells, Marcus Tutt, & Sanja Abbott. "Tessa, a system to aid communication with deaf people." In *Proceedings of the 5th international ACM conference on assistive technologies*, pp. 205–212, 2002.

[48] Inge Zwitserlood, Margriet Verlinden, Johan Ros, & Sanny van der Schoot. "Synthetic signing for the deaf: esign." Visicast, pp. 1–6, 2005.

[49] Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, & François Lefebvre-Albaret. "The dicta-sign wiki: Enabling web communication for the deaf." *International Conference on Computers for Handicapped Persons (ICCHP)*, pp. 205–212, 2012.

[50] Virtual Humans Group. "Virtual humans research for SL animation." School of Computing Sciences, UEA Norwich, UK, 2017.

[51] Mohammed Jemni, Oussama El Ghoul, Mehrez Boulares, Nour Ben Yahia, Kabil Jaballah, Achraf Othman, & Monoem Youneb. "Websign." Latice, 2020.

[52] Siegmund Prillwitz. *Hamnosys. version 2.0. Hamburg notation system for SLs. An introductory guide.* Hamburg Signum Press, 1989.

[53] Richard Kennaway. "Avatar-independent scripting for real-time gesture animation." *Procedural animation of SL*, arXiv:1502.02961, 2013.

[54] Masahiro Mori, Karl F. MacDorman, & Norri Kageki. "The uncanny valley [from the field]." *IEEE Robotics & Automation Magazine*, vol. 19, pp. 98–100, 2012.

[55] Sarah Ebling & Matt Huenerfauth. "Bridging the gap between SL machine translation & SL animation using sequence classification." *Proceedings of SLPAT 2015: 6th Workshop on Speech & Language Processing for Assistive Technologies*, pp. 2–9, 2015.

[56] Sarah Ebling & John Glauert. "Exploiting the full potential of jasigning to build an avatar signing train announcements." In *3rd International symposium on SL translation & avatar technology*, pp. 1–9, 2013.

[57] Sylvie Gibet, François Lefebvre-Albaret, Ludovic Hamon, Rémi Brun, & Ahmed Turki. "Interactive editing in French SL dedicated to virtual signers: Requirements & challenges." *Universal Access in the Information Society*, vol. 15, pp. 525–539, 2016.

[58] Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, & Richard Bowden. "Text2sign: Towards SL production using neural machine translation & generative adversarial networks." *International Journal of Computer Vision*, vol. 128, pp. 891–908, 2020.

[59] Achraf Othman & Mohamed Jemni. "Statistical SL machine translation: from English written text to American SL gloss." *IJCSI International Journal of Computer Science*, vol. 8, pp. 65–73, 2011.

[60] Minh-Thang Luong, Hieu Pham, & Christopher D. Manning. "Effective approaches to attention-based neural machine translation." arXiv:1508.04025, 2015.

[61] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, & Richard Bowden. "Neural SL translation." CVPR, 2018.

[62] Dan Guo, Wengang Zhou, Houqiang Li, & Meng Wang. "Hierarchical LSTM for SL translation." *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

[63] Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, & Koray Kavukcuoglu. "Neural machine translation in linear time." arXiv:1610.10099, 2016.

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, & Illia Polosukhin. "Attention is all you need." NIPS, 2017.

[65] Jehee Lee & Sung Yong Shin. "A hierarchical approach to interactive motion editing for human-like figures." SIGGRAPH ’99: Proceedings of the 26th annual conference on Computer graphics & interactive techniques, pp. 39–48, 1999.

[66] Qifeng Chen & Vladlen Koltun. "Photographic image synthesis with cascaded refinement networks." ICCV, pp. 1511–1520, 2017.

[67] Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, & Koray Kavukcuoglu. "Conditional image generation with pixelcnn decoders." NIPS, 2016.

[68] Aaron van den Oord, Nal Kalchbrenner, & Koray Kavukcuoglu. "Pixel Recurrent Neural Networks." In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1747–1756, 2016.

[69] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, & Daan Wierstra. "DRAW: A Recurrent Neural Network for Image Generation." In *Proceedings of Machine Learning Research*, 2015.

[70] Diederik P. Kingma & Max Welling. "Auto-Encoding Variational Bayes." In *International Conference on Learning Representations (ICLR)*, 2014.

[71] Xinchen Yan, Jimei Yang, Kihyuk Sohn, & Honglak Lee. "Attribute2Image: Conditional Image Generation from Visual Attributes." In *European Conference on Computer Vision (ECCV)*, pages 776–791, 2016.

[72] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, & Yoshua Bengio. "Generative Adversarial Nets." In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[73] Aliaksandr Siarohin, Enver Sangineto, Stephane Lathuiliere, & Nicu Sebe. "Deformable GANs for Pose-Based Human Image Generation." In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, 2018.

[74] Yuusuke Kataoka, Takashi Matsubara, & Kuniaki Uehara. "Image Generation Using Adversarial Networks & Attention Mechanism." In *IEEE/ACIS 15th International Conference on Computer & Information Science (ICIS)*, 2016.

[75] Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, & Koray Kavukcuoglu. "Conditional Image Generation with PixelCNN Decoders." In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[76] Jan Zelinka & Jakub Kanis. "Neural SL Synthesis: Words Are Our Glosses." In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3395–3403, 2020.

[77] Ben Saunders, Necati Cihan Camgöz, & Richard Bowden. "Everybody Sign Now: Translating Spoken Language to Photorealistic SL Video." *arXiv preprint arXiv:2011.09846*, 2020.

[78] Sandrine Tornay, Necati Cihan Camgoz, Richard Bowden, & Magimai Doss. "A Phonology-Based Approach for Isolated Sign Production Assessment in SL." In *Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion)*, 2020.

[79] Michael Kipp, Alexis Heloir, & Quan Nguyen. "SL Avatars: Animation & Comprehensibility." In *International Workshop on Intelligent Virtual Agents*, pages 113–126, 2011.

[80] John McDonald, Rosalee Wolfe, Jerry Schnepp, Julie Hochgesang, Diana Gorman Jamrozik, Marie Stumbo, Larwan Berke, Melissa Bialek, & Farah Thomas. "An Automated Technique for Real-Time Production of Lifelike Animations of American SL." *Universal Access in the Information Society*, 15:551–566, 2016.

[81] Ben Saunders, Necati Cihan Camgöz, & Richard Bowden. "Adversarial Training for Multi-Channel SL Production." In *British Machine Vision Conference (BMVC)*, 2020.

[82] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, & Philipp Slusallek. "Synthesis of Compositional Animations from Textual Descriptions." In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1396–1406, 2021.

[83] Wencan Huang, Wenwen Pan, Zhou Zhao, & Qi Tian. "Towards Fast & High-Quality SL Production." In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3172–3181, 2021.

[84] Joseph B. Kruskal. "An Overview of Sequence Comparison: Time Warps, String Edits, & Macromolecules." *SIAM Review*, 25(2):201–237, 1983.

[85] Ozge Mercanoglu Sincan & Hacer Yalim Keles. "AUTSL: A Large Scale Multi-Modal Turkish SL Dataset & Baseline Methods." *IEEE Access*, 8:181340–181355, 2020.

[86] Mujahid A., Awan M. J., Yasin A., Mohammed M. A., Damaševičius R., Maskeliūnas R., Abdulkareem K. H. "Real-Time Hand Gesture Recognition Based on Deep Learning YOLOv3 Model." *Applied Sciences*, 11(9):4164, 2021.

[87] Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, & Richard Bowden. "Text2Sign: Towards SL Production Using Neural Machine Translation & Generative Adversarial Networks." *International Journal of Computer Vision*, 128, 891–908, 2020.

[88] Vasileios Baltatzis, Necati Cihan Camgoz, Ben Saunders, Stavros Petridis, Maja Pantic, Richard Bowden, & Philip Torr. "Neural Sign Actors: A Diffusion Model for 3D SL Production from Text." In *Proceedings of the IEEE/CVF Conference on Computer Vision & Pattern Recognition (CVPR)*, 2024.

[89] Shalev-Arkushin, Amit Moryossef & Ohad Fried. "Ham2Pose: Animating SL Notation into Pose Sequences." In *Proceedings of the IEEE/CVF Conference on Computer Vision & Pattern Recognition (CVPR)*, 2023.

[90] Razieh Rastgoo and Kourosh Kiani and Sergio Escalera and Mohammad Sabokrou. “Sign Language Production: A Review”