



**FPT UNIVERSITY**

# **SIGN LANGUAGE TRANSLATION SYSTEM**

**Chi Ho**

**Supervisors:** Trung Nguyen Quoc, Vinh Truong Hoang

**DEPARTMENT OF ITS**

**FPT UNIVERSITY HO CHI MINH**

A final year capstone project submitted in partial fulfillment of the requirement  
for the Degree of Bachelor of Artificial Intelligent in Computer Science

April 2024

# ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisors, Professor Trung Nguyen Quoc and Professor Vinh Truong Hoang, for their invaluable guidance and support throughout the duration of this thesis. Their invaluable guidance, drawn from their wealth of experience and profound expertise, significantly enhanced the completion of this study. My sincere thanks also goes to the Perception Technology Solution company for providing conducive conditions for my study, and to my leader, Mr. Khang, for his mentorship in executing this research endeavor.

I extend my appreciation to Professor Tai Nguyen Trong and Professor Hai Le Thanh for their constructive feedback during the review process, which greatly contributed to enhancing the quality of this thesis. Additionally, I wish to express special thanks to Ms. Nghia from the Lam Dong School for the Deaf for generously sharing her insights into sign language and deaf culture. Despite the distinctions between Vietnamese Sign Language and American Sign Language, her contributions have profoundly enriched my understanding and research endeavors in this domain.

# AUTHOR CONTRIBUTIONS

This thesis, including the research, data preparation, method exploration, training, evaluation, experimentation, analysis, demo software, documentation and report writing, was conducted solely by Chi Ho. As the sole contributor, Chi Ho takes full responsibility for all aspects of this work.

# ABSTRACT

Sign language translation systems play a crucial role in eliminating communication barriers between Deaf and Hard-of-Hearing (DHH) individuals and those who can hear. These systems are complex and consist of multiple components. Among them, the automatic translation of spoken language into sign language, known as Sign Language Production (SLP), holds the potential to revolutionize sign language communication applications. Contrary to its importance and necessity, research in SLP still lacks depth, with only a few models publicly available with insufficient evaluations and comparisons. Consequently, the comprehension of SLP experiments remains obscure, with few new studies emerging in this domain. This project endeavors to investigate the efficacy of existing public SLP methods on American Sign Language (ASL). Specifically, the experiment involved training and evaluating three distinct approaches (Regressive Training with Progressive Transformers, Adversarial Training with Progressive Transformers, and Non-Autoregressive Transformers with Gaussian Space) using the How2Sign dataset, one of the most comprehensive datasets comprising instructional videos in American Sign Language. Back-translation evaluation metrics were employed to assess the performance of these methods in translating discrete spoken language sentences into continuous 3D sign pose sequences. The results indicate that, for the complex data involved in translating from spoken language to sign language in SLP, Non-Autoregressive Transformers with Gaussian Space (NSLP-G) outperform other methods, accurately capturing both manual and non-manual features with minimal errors. Additionally, with the Progressive Transformers model, the effectiveness of adversarial compared to sole regressive training in translate from text to sign language field is observed. Within the scope of this thesis project, a Minimum Viable Product (MVP) was developed to test real-time text-to-sign language translation. The project's outcomes can provide valuable insights for future researchers, guiding them towards viable approaches in exploring this field or considering practical applications. The report also outlines the limitations of this project and proposes future work that could be utilized to further develop and improve sign language production models.

**Keywords:** *Sign Language Production, Continuous Sign Language Generation, Progressive Transformers, Adversarial Training, Non-Autoregressive Transformer, Variational Autoencoder, Human Pose Generation.*

# CONTENTS

<b>1</b>	<b>INTRODUCTION</b>	<b>9</b>
1.1	Overview . . . . .	10
1.2	Motivation . . . . .	10
1.2.1	Practical motivation . . . . .	10
1.2.2	Technical motivation . . . . .	11
1.3	Project Objectives and Chapter Overview . . . . .	11
<b>2</b>	<b>BACKGROUND</b>	<b>14</b>
2.1	Sign Language Linguistics . . . . .	14
2.1.1	Phonology . . . . .	14
2.1.2	Simultaneity . . . . .	14
2.1.3	Referencing . . . . .	15
2.1.4	Fingerspelling . . . . .	15
2.2	Sign Language Representations . . . . .	15
2.2.1	Videos . . . . .	15
2.2.2	Skeletal Poses . . . . .	16
2.2.3	Written notation systems . . . . .	16
2.2.4	Gloss . . . . .	16
<b>3</b>	<b>RELATED WORKS</b>	<b>17</b>
3.1	Sign Language Processing . . . . .	17
3.2	Sign Language Production . . . . .	17
3.3	Sign Language Dataset . . . . .	18
<b>4</b>	<b>PROJECT MANAGEMENT PLAN</b>	<b>19</b>
4.1	Overview . . . . .	19
4.2	Project Scope and Objectives . . . . .	19
4.3	Project Schedule . . . . .	20
<b>5</b>	<b>MATERIALS AND METHODS</b>	<b>21</b>
5.1	Data . . . . .	21
5.1.1	Data Preprocessing . . . . .	22
5.2	Methods . . . . .	23
5.2.1	Regressive Training with Progressive Transformers . . . . .	23
5.2.2	Adversarial Training with Progressive Transformers . . . . .	27
5.2.3	Non-Autoregressive Transformers with Gaussian Space . . . . .	30
5.3	Implementation Setup . . . . .	34

---

5.4	Metrics . . . . .	34
<b>6</b>	<b>RESULTS</b>	<b>35</b>
6.1	Quantitative Results . . . . .	35
6.2	Qualitative Results . . . . .	35
<b>7</b>	<b>DISCUSSIONS</b>	<b>38</b>
7.1	Interpretation and Implications . . . . .	38
7.2	Limitations and Future Works . . . . .	39
<b>8</b>	<b>CONCLUSIONS</b>	<b>40</b>
<b>A</b>	<b>Demo</b>	<b>41</b>
A.1	UI Structure . . . . .	41
A.2	Use cases . . . . .	42
A.2.1	Get a random sample of text and its sign language video . . . . .	42
A.2.2	Duplicate the text to translate panel input . . . . .	42
A.2.3	Translate the text to sign language pose . . . . .	42
A.2.4	Download the sign language pose video . . . . .	43

# LIST OF FIGURES

1	<b>Data preprocessing pipeline.</b> Left to right: original image, original pose extracted by Mediapipe Holistic, pose removed unnecessary landmarks, pose aligned hand wrists with body wrists, and final pose after transferring appearance. . . . .	22
2	<b>Pose anonymization procedure.</b> Left to right: first frame, the reference appearance, pose brought back the hands, pose brought back the wrists. .	23
3	<b>Architecture details of Progressive Transformers.</b> (PE: Positional Encoding, CE: Counter Embedding, MHA: Multi-Head Attention). . . .	25
4	<b>An illustration of Counter Decoding.</b> Demonstrate the concurrent prediction of the sign pose, $\hat{y}_u$ , and the counter, $\hat{c}_u$ , where $\hat{c}_u \in 0 : 1$ , and $\hat{c}_u = 1.0$ denotes the end of the sequence. . . . .	26
5	<b>An overview of Adversarial Multi-Channel SLP.</b> Featuring a Conditional Adversarial Discriminator assessing the realism of Sign Pose Sequences generated by an SLP Generator. . . . .	28
6	<b>Details of our Conditional Adversarial Discriminator architecture.</b> The sign pose sequence $Y_{1:U}$ is concatenated with the source spoken language sequence $X_{1:T}$ and mapped to a single scalar, $d_p$ . . . . .	29
7	<b>An overview of the NSLP-G model.</b> Comprising two phases for generating sign poses based on a given source sentence. Phase I: VAE is utilized to serve as Gaussian Pose Generator (GPG). The encoder produces $\mu$ and $\sigma$ , and then employs a reparameterization trick to sample $z$ following $N(0, 1)$ . The decoder reconstructs $\hat{y}$ using $z$ . Phase II: To generate a sequence of $z$ , a Transformer-equipped GPG is employed to produce a sequence of $z$ . To achieve non-autoregressiveness, autoregressive connections are eliminated from the decoder, and only positional encodings (PE) are utilized as inputs. . . . .	31
8	<b>A depiction of the Transformer-based Gaussian Seeker.</b> This module comprises a transformer encoder and a non-autoregressive decoder. The encoder processes the source sentence $x_0, \dots, x_t, \dots, x_T$ consisting of $U$ words, while the decoder takes positional encodings (PE) with a length of $U$ as input to generate a sequence of latent vectors $Z$ following a Gaussian distribution. . . . .	33
9	<b>Visualization of generated sign pose sequences in Case 1 evaluation.</b> Observations indicate varying levels of accuracy and deviation from the original poses across the models, highlighting the influence of autoregressive and non-autoregressive approaches on pose generation. . .	37

---

10	<b>Visualization of generated sign pose sequences in Case 2 evaluation.</b> Despite limitations in visualizing poses, certain non-manual features such as facial expressions and head movements are partially evident in the generated sequences. NSLP-G exhibits the closest resemblance to the original non-manual features among the evaluated models. . . . .	37
11	<b>The demo UI structure.</b> The UI includes 2 main parts. . . . .	41
12	<b>Use case with Random button.</b> Users click on Random button. . . .	42
13	<b>Use case with Random button.</b> Result of user clicking the random button. . . . .	42
14	<b>Use case with Duplicate button.</b> Users click Duplicate button. . . .	43
15	<b>Use case with Duplicate button.</b> Result of users clicking the Duplicate button. . . . .	43
16	<b>Use case with Translate button.</b> Users click the Translate button. . .	43
17	<b>Use case with Translate button.</b> Result of users clicking the Translate button. . . . .	44
18	<b>Use case with Download button.</b> Users click the Download button. .	44
19	<b>Use case with Download button.</b> Result of uses clicking the Download button. . . . .	44



# LIST OF TABLES

1	<b>Project information.</b> Supervisor information . . . . .	19
2	<b>Project information.</b> Student information . . . . .	19
3	<b>Project schedule.</b> Overview of the timeline for the Sign Language Translation System thesis . . . . .	20
4	<b>How2Sign Dataset Overview.</b> Statistics of the original and final versions of the How2Sign dataset. . . . .	21
5	<b>Performance comparison of three approaches to Text-to-Pose task.</b> Results indicate that the Adversarial Training regime significantly enhances performance compared to the Progressive Transformers trained solely with a Regression loss. The Non-Autoregressive approach, particularly NSLP-G, demonstrates the highest performance, showcasing a distinct advantage over the other methods, which operate in an Autoregressive manner. . . . .	36

## Chapter 1

# INTRODUCTION

Since the dawn of humanity, communication has played a pivotal role in the development as a species. Today, it's hard to imagine living in a world where others cannot understand one another, individuals struggle to comprehend what others say. This disconnect in communication profoundly affects both individual and societal development. However, these barriers are precisely what Deaf and Hard-of-Hearing (DHH) individuals face. Their communication language, Sign Languages (SLs), utilizes visual-gestural modalities to convey meaning through manual articulations combined with non-manual elements like facial expressions and body movements. While two Deaf individuals without physical disabilities can communicate with each other using sign language, communication becomes significantly challenging when a Hearing person interacts with a Deaf or Hard-of-Hearing individual, and vice versa. Sign language is difficult to learn and inaccessible to those outside the Deaf community. Moreover, due to audism, DHH individuals are often disregarded or compelled to use alternative communication methods with which they may not be comfortable (e.g. write down any message or use special gloves for the detection of signs), creating further barriers to daily communication.

In the context of the rapidly evolving field of AI, particularly with notable successes in machine translation, the development of models for translating between spoken language and sign language becomes more feasible than ever before. This advancement is crucial for the Deaf community because translation provides a communication bridge between DHH individuals and the Hearing population, thereby granting DHH individuals access to information like everyone else.

This Introduction chapter lays the groundwork for exploring the subsequent chapters, providing a comprehensive overview of the purpose, objectives, and research methods used in this study. Following this introduction, the subsequent sections will provide an overview of the project, followed by an exploration of what motivated the researchers to conduct this study both practically and technically. Subsequently, specific objectives will be outlined, and the structure of this report will be briefly described to lead into the following sections.

## 1.1 Overview

Sign languages serve as vital communicative tools within the deaf and hard-of-hearing (DHH) community and are a central element of Deaf culture. The World Health Organization (WHO) reports that over 5% of the world's population, totaling 430 million people, require rehabilitation to address their disabling hearing loss, which includes 34 million children. It is estimated that by 2050, this number will surpass 700 million people, equating to 1 in every 10 individuals worldwide experiencing disabling hearing loss [1]. According to National Geographic, there are 300 different forms of sign language around the world. [2]. These figures underscore the significant portion of the population in need of alternative communication methods, especially considering the predominance of an aural society that often marginalizes DHH individuals, leading to audism and isolation.

Sign languages offer a solution to bridge communication gaps between signers, removing barriers posed by verbal languages. However, the challenge persists when individuals cannot communicate using signs, creating a substantial barrier between the hearing and DHH communities. Sign languages, utilizing the visual-gestural modality, are recognized as natural languages, complete with their own grammar and lexicons [3]. Proficiency in sign language among hearing individuals remains limited, often resulting in situations where Deaf and Hard-of-Hearing (DHH) individuals resort to alternative communication methods, such as writing messages or utilizing specialized gloves for sign detection. For Deaf individuals, sign language is not merely a tool but a fundamental aspect of their identity and culture, offering a more natural and comprehensive means of communication compared to written forms of spoken language. To address this, non-intrusive communication tools adaptable to both signers and non-signers must be developed, ensuring the comfort and inclusivity of the DHH community.

Developing a robust system for translating spoken languages into sign languages and vice versa is essential to bridging communication gaps between the DHH and hearing communities. This translation system comprises two tasks: Sign Language Translation (SLT) functions to translate from sign to text, while Spoken Language Translation (SLP) translates text to sign. This project focuses specifically on the SLP task, translating spoken language into sign language. The following section 1.2, Motivation, will further elucidate why Sign Language Production was chosen for this project in both practical and technical motivation.

## 1.2 Motivation

### 1.2.1 Practical motivation

Throughout history, Deaf communities have advocated for the recognition and use of sign languages, asserting their legitimacy as sophisticated communication modalities on par with spoken languages in all linguistic and social aspects [3]. However, in predominantly oral societies, deaf individuals are often encouraged to rely on spoken languages, either through lip-reading or text-based communication. Previous researchs have primarily focused on Sign Language Translation (SLT), translates from sign language into spoken language, catering mainly to hearing individuals who can receive information through their native language (text/speech). However, this approach does not adequately

address the needs of Deaf and Hard-of-Hearing (DHH) individuals, whose natural mode of communication is sign language. Deaf communities strongly prefer to communicate in sign languages, both online and in face-to-face interactions, both among themselves and with spoken language communities [4, 5]. Therefore, creating translation systems proficient in converting spoken language into sign language is crucial for eliminating language barriers, enabling individuals to communicate more comfortably and naturally with one another.

Furthermore, Sign Language Production (SLP) is a critical research domain with significant potential to positively impact signing communities. Sign language technologies offer various applications such as documenting endangered sign languages, providing educational tools for sign language learners, enabling information retrieval from sign language videos, developing personal assistants responsive to sign languages, offering real-time automatic sign language interpretations, and more.

### 1.2.2 Technical motivation

While research on translating Sign languages into spoken languages has made significant strides in recent years [6, 7, 8, 9, 10, 11], the translation of spoken languages into Sign languages, known as Sign Language Production (SLP), remains a formidable challenge [12, 13, 14, 15]. One reason for the limited progress in SLP is the misconception that deaf individuals are proficient in reading spoken language and do not require translation into Sign language. Another challenge is the scarcity of publicly available methods serving as baselines for further research, making synthesis, evaluation, and comparison difficult for subsequent studies to build upon and extend. Consequently, current SLP research is still in its infancy, with few publicly accessible models and evaluation criteria.

From a technical perspective, Sign Language Production (SLP) presents a unique array of challenges due to the complexity and richness of sign languages. This task involves both Natural Language Processing (NLP) and Computer Vision (CV) techniques. Unlike spoken languages, sign languages rely on visual-gestural modalities, necessitating specialized techniques for translation.

This project aims to confront these technical challenges by investigating the efficacy of various public SLP approaches, with the goal of advancing the state-of-the-art in SLP technology and contributing to the development of more accurate and efficient translation systems for sign languages.

## 1.3 Project Objectives and Chapter Overview

The project aims to experiment with publicly available models for the Sign Language Production (SLP) task to evaluate and compare their performance. The goal is to provide essential insights for future research and development efforts, addressing the current limitations of existing models. Additionally, the project serves as a foundation for determining the most effective approach to apply in real-world sign language translation systems.

Following an in-depth exploration, three distinct approaches will be experimented with, including:

- Approach 1: Regressive Training with Progressive Transformers [12], which is the first SLP model capable of translating from text to continuous 3D sign pose sequences in an end-to-end manner.
- Approach 2: Adversarial Training with Progressive Transformers [16], which employs a progressive transformer architecture with a conditional adversarial discriminator, supplementing the regression loss with an adversarial loss.
- Approach 3: Non-Autoregressive Transformers with Gaussian Space [17], where the non-autoregressive Transformer translates the source sentence to the target sign pose distributions based on Variational Autoencoder in Gaussian space.

These three methods were chosen because they represent different approaches to the problem, which, while familiar in conventional machine translation tasks, are rarely directly compared in sign language translation research. Furthermore, there is a scarcity of public models for SLP, and these three methods are among the few with published code, often serving as benchmarks for newer approaches.

It's noteworthy that sign language translation typically involves an intermediary notation system such as Gloss, HamNoSys, or SignWriting, ... (mentioned in Section 2.2) and the aforementioned approaches offer options to translate through or combine with Gloss data to enhance performance. However, in this project, the models are trained and evaluated directly without using Gloss notation. This decision stems from the observation that end-to-end translation yields better performance compared to approaches involving intermediary sub-tasks, such as translating Text to Gloss and then Gloss to Pose. This is attributed to the richness of information available in spoken language compared to Gloss representations, which may act as a bottleneck, potentially leading to the loss of contextual information from the source text [12]. Moreover, datasets with Gloss annotation are scarce, costly to label, and require specialized knowledge of sign language, making them less practical.

It's worth noting that in this research, sign language is represented in the form of 3D sign pose sequences, advantageous for direct video generation as the final animation can be performed by an ordinary avatar. This approach not only enhances training results by reducing data dimensions but also facilitates real-world applications, including real-time scenarios. Additionally, output sign poses serve well for precise and specific research purposes.

All poses in the research encompass both manual and non-manual features, as non-manual features are crucial for understanding sign language, providing grammatical syntax, context, and emphasis. Saunders and his colleagues has demonstrated significantly higher performance when training with both manual and non-manual features compared to training with manual features alone [16].

Furthermore, the project focuses on researching and evaluating models for American Sign Language (ASL), which is the predominant sign language of deaf communities in the United States and Anglophone Canada. ASL is a complete visual language expressed through both manual and non-manual features. Therefore, all evaluation studies in this project are conducted on the How2Sign dataset, a high-quality ASL dataset that meets the project's requirements.

In summary, the project conducts research and comparison of three different approaches to the SLP problem on American Sign Language datasets. Specifically, the project's contributions include:

1. Experimentation with three different SLP approaches on the How2Sign dataset, including regressive training, adversarial training, and non-autoregressive modeling.
2. Comparison and analysis of the performance of these SLP approaches.
3. Provision of insights and recommendations for future research and development efforts in the field of SLP, aiming to enhance communication accessibility for individuals who use sign language.
4. Development of a sign language translation system in the form of a demo.

The structure of this report is as follows: Chapter 2 offers background information on sign language. In Chapter 3, we delve into existing literature on sign language processing, production, and datasets. Chapter 4 provides an overview of the project. Details regarding data preprocessing, the architecture of the three approaches, implementation setup, and metric evaluation are covered in Chapter 5. Chapter 6 presents both quantitative and qualitative results. In Chapter 7, we discuss these findings, offer insights, limitations and propose recommendations for future research. Finally, Chapter 8 wraps up the report. Additionally, Appendix A contains information of the demo for this thesis project.

## Chapter 2

# BACKGROUND

The Background chapter serves as the foundational framework upon which the entirety of this thesis is built. Its primary aim is to provide a comprehensive understanding of sign language, encompassing both its linguistic characteristics and modes of representation. By delving into the intricate components of sign language linguistics and its various forms of representation, this chapter sets the stage for a deeper exploration of the research topic.

## 2.1 Sign Language Linguistics

Signed languages are structured similarly to spoken languages, with elements like morphological, phonological, syntactic, and semantic structures. Instead of talking, people use their hands, face, and body to communicate. This paper will highlight the linguistic attributes of signed languages that researchers need to incorporate into their models.

### 2.1.1 Phonology

Signs are made up of basic components that include manual aspects like hand shape, palm direction, location, contact, movement path, and localized movement, as well as non-manual aspects like eye openness, head motion, and body positioning [18, 19, 20, 21]. Both signed and spoken languages don't always use all possible sounds, and the sets of sounds or features in two languages might not match entirely. Each language also follows rules about how these features can be combined.

### 2.1.2 Simultaneity

Despite taking approximately twice as long to produce compared to an English word, an ASL sign transmits information at a similar rate to spoken English [22]. Signed languages address the slower production pace of signs through simultaneity, utilizing multiple visual cues simultaneously to convey various information [21]. For instance, a signer might sign "cup" with one hand while pointing to the actual cup with the other to indicate "that cup." Analogous to intonation in spoken languages, facial expressions and body posture convey additional emotional nuances [23, 24]. Facial gestures can alter adjectives, adverbs, and verbs; a nod can negate a phrase or sentence; eye gaze can help identify referents.

### 2.1.3 Referencing

In conversation, signers can introduce subjects either by pointing to their actual positions in space or by designating a specific area in the signing space to represent a subject not physically present, and then indicating that area to refer to it [25, 26]. Furthermore, signers establish connections between subjects anchored in the signing space by using directional signs or physically embodying the subjects through shifts in body position or eye gaze [27, 28]. Spatial references also influence grammatical structure, as the direction of a verb may depend on the placement of the subject and/or object references [29, 30]. For example, a directional verb might originate from the location of the subject and conclude at the location of the object. While the relationship between subjects and verbs in spoken language tends to be arbitrary, in signed languages, subject relationships are typically rooted in spatial context. Visual space is extensively utilized to ensure clarity in referencing.

In sign language, classifiers or depicting signs [31, 32, 33] are used to describe referents. These signs, often one-handed and flexible in movement, convey details about the referent's characteristics, movement, and relationships with other entities [23]. For instance, to describe a car swerving and crashing, a signer might use a hand classifier for a vehicle, gesture swerving, and simulate a crash with another entity in space.

Signers employ role shift [34] to quote others, physically embodying their characteristics. For example, in recounting a dialogue between a tall and a short person, the signer may shift position and adjust their gaze accordingly.

### 2.1.4 Fingerspelling

Fingerspelling comes from how signed languages interact with written forms of spoken languages [35, 36, 37]. It involves using hand movements to spell out words or letters. While it's often used for names, places, or new ideas from spoken language, it's also become a regular part of signed languages [38, 39].

## 2.2 Sign Language Representations

Representing signed languages poses a major hurdle for SLPs. Unlike spoken languages, signed languages lack a widely accepted written form. Since signed languages rely on visual and gestural communication, video recording is the most direct method to document them. However, videos contain excessive information for modeling and are costly to produce, store, and share. Hence, researchers have been striving to find a more efficient, lower-dimensional representation.

### 2.2.1 Videos

Videos are the most direct way to represent signed languages, effectively capturing the information conveyed through signing. However, one significant drawback is their high dimensionality: Videos often contain more data than necessary for modeling, making them costly to store, transmit, and encode. Anonymizing raw videos, crucial for protecting privacy, remains a challenge, hindering their widespread public availability [40].



## 2.2.2 Skeletal Poses

Videos can be simplified into skeleton-like wireframes or meshes, showing joint locations. This helps estimate human pose from video data, determining body configuration over time. While motion capture equipment offers high-quality results, it's costly and invasive. As a result, pose estimation from videos has become more popular [41, 42, 43, 44]. Skeletal poses offer a less complex and partially anonymized representation of the body, with minimal information loss. However, they're not well-suited for most NLP models due to their continuous, multidimensional nature.

## 2.2.3 Written notation systems

Different systems represent signs visually, with some using linear writing and others employing two-dimensional graphemes. Despite various proposed universal [45, 46] and language-specific notation systems [47, 48, 49] no single writing system has been widely adopted by any sign language community. This lack of standardization makes it challenging to share and integrate resources across projects. The figure above shows two universal notation systems: SignWriting [45], a two-dimensional pictographic system, and HamNoSys [46], a linear sequence of graphemes designed for machine readability.

## 2.2.4 Gloss

Glosses transcribe signed languages sign-by-sign, assigning each sign a unique semantic identifier. While several sign language corpus projects offer guidelines for gloss annotation [50, 51, 52], a standardized protocol for gloss annotation is yet to be established. Linear gloss annotations have been criticized for their imprecise representation of signed language, as they fail to capture the simultaneous expression of information through various cues like body posture, eye gaze, or spatial relations. This loss of information can significantly impact the performance of downstream tasks in SLP [53].

## Chapter 3

# RELATED WORKS

The Related Works chapter offers a thorough examination of current research and advancements in sign language processing. It delves into key areas such as sign language recognition, translation, and production, particularly emphasizing the translation of video-based sign language into text-based sequences. Additionally, recent progress has been made in addressing sign language datasets. The objective of this chapter is to scrutinize and assess the methodologies and datasets utilized in prior research endeavors.

### 3.1 Sign Language Processing

Sign language processing encompasses various research directions, including sign language recognition (SLR), sign language translation (SLT), sign spotting [6, 54, 55, 56], and sign language retrieval [57, 58]. SLR aims to transcribe sign videos into their constituent glosses, which can be categorized into isolated SLR (ISLR) [59, ?, 60, 61, 62] and continuous SLR (CSLR) [63, 64, 65]. ISLR focuses on predicting the gloss of an isolated sign, while CSLR recognizes sequences of signs in videos and generates corresponding gloss sequences. SLT goes a step further by translating sign languages into spoken languages. Recent works have formulated SLT as a neural machine translation problem [66, 8, 63, 67], utilizing visual encoders and language models. Inspired by the success of transferring pre-trained language models to SLT [64], mBART [68] is adopted as the Text2Gloss translator in this study.

### 3.2 Sign Language Production

Stoll et al. [14, 69] introduced the first deep SLP model, employing a three-step pipeline. Initially, they established a mapping between sign glosses and skeleton poses via a lookup table for Grapheme-to-Phoneme (G2P) conversion. Building upon this, B. Saunders et al. [12] proposed the progressive transformer, utilizing an encoder-decoder architecture to learn the mapping and generate sign poses in an autoregressive manner during inference. Subsequently, B. Saunders et al. [70] introduced a Mixture Density Network (MDN) to generate pose sequences conditioned on sign glosses, leveraging a GAN-based method [71] to produce photo-realistic sign language videos. They further enhanced spoken language to sign language translation using an autoregressive transformer network and incorporating gloss information for additional supervision [13], or

applied adversarial learning [16]. Additionally, they proposed a Mixture of Motion Primitives (MoMP) architecture to combine distinct motion primitives for continuous sign language sequence production. In a subsequent work, B. Saunders et al. [72] introduced a Frame Selection Network (FS-NET) to improve temporal alignment and SIGNGAN, a pose-conditioned human synthesis model for generating photo-realistic sign language videos directly from skeleton poses. Despite achieving state-of-the-art results, they relied on an additional sign language dictionary [73], limiting direct comparison of their findings. In contrast, Huang et al. [74, 17] proposed a non-autoregressive model to simultaneously generate sign pose sequences, addressing the error accumulation issue by employing monotonic alignment search to determine the alignment lengths of each gloss.

### 3.3 Sign Language Dataset

A significant obstacle hindering progress in sign language research has been the scarcity of large-scale datasets [75]. Early datasets were primarily focused on isolated sign recognition, containing only a limited vocabulary [76, 61, 62, 77, 78]. However, to address the challenges of sign language recognition and translation in the context of complete sentences, several continuous sign language datasets have been developed. Examples include RWTH-BOSTON-50 [79], Dreuw et al. [80], SIGNUM [81], BSL [82], and the DictaSign Corpus, which provided sentence-level annotations in multiple languages [83, 84, 85]. Despite the introduction of additional datasets featuring expanded sign sets [86, 61, 62], the importance of continuous sign language for translation and production cannot be overstated.

Among the pioneering continuous sign language datasets is S-pot [87], which offered over 1,000 signs of Finnish sign language in a controlled environment. Another significant dataset, RWTH-Phoenix [7], comprised TV clips with German sign language and remains widely used for sign language translation and production. Similarly, BSL-1K [8] assembled a collection of British sign language signs used in everyday conversations, totaling 1,000 signs. More recently, Duarte et al. introduced How2Sign [88], a large-scale dataset of American Sign Language (ASL) aligned with speech signals from the How2 dataset. How2Sign boasts a vocabulary of over 16,000 signs, captured over seventy-nine hours of continuous sign language.

## Chapter 4

# PROJECT MANAGEMENT PLAN

The Project Management Plan chapter serves as a comprehensive roadmap for the successful execution of this thesis project. Within this chapter, an overview of the project will be provided, followed by detailed information regarding the project's objectives, scope, and schedule.

### 4.1 Overview

This is the graduation project of a student majoring in Artificial Intelligence at FPT University, Spring 2024 semester. The project information includes:

- Information about supervisors:

	Full name	Email	Title
Supervisor 1	Nguyen Quoc Trung	trungnq46@fe.edu.vn	Mr.
Supervisor 2	Truong Hoang Vinh	vinhth8@fe.edu.vn	Dr.

Table 1. **Project information.** Supervisor information

- Information about the project team:

	Full name	Student ID	Email	Role in Group
Student 1	Ho Linh Chi	SE150666	chihlse150666@fpt.edu.vn	Leader

Table 2. **Project information.** Student information

### 4.2 Project Scope and Objectives

The objective of this project is to assess and analyze current Sign Language Production (SLP) methodologies, particularly in the context of American Sign Language (ASL). Three distinct SLP models were selected for experimentation, each representing a different approach. These models were trained and evaluated using the How2Sign dataset. Additionally, a Minimum Viable Product (MVP) was developed as part of this graduation project to enable real-time text-to-sign language translation, as outlined in Appendix A.

Through the fulfillment of these objectives, the project aims to contribute to the progression of SLP technology, ultimately fostering greater inclusivity in communication for the Deaf and Hard-of-Hearing communities.

### 4.3 Project Schedule

Table 3 provides an overview of the timeline for this project. The schedule outlines the estimated durations for each phase, which may be subject to adjustments based on project progress and requirements.

Task name	Priority	Start date	End date	Status
Find documents	High	01/01/2024	07/01/2024	Done
Review papers	Medium	05/01/2024	14/01/2024	Done
Review and analyze public dataset	Low	15/01/2024	21/01/2024	Done
Research on models and architectures	High	22/01/2024	04/02/2024	Done
Find and choose 3 different approaches	High	05/02/2024	11/02/2024	Done
Experiment models	High	13/02/2024	10/03/2024	Done
Compare results	High	11/03/2024	17/03/2024	Done
Write report	High	18/03/2024	15/04/2024	Done
Review Report	Medium	15/04/2024	18/04/2024	Done
Create Web demo	Low	18/03/2024	31/03/2024	Done
Revision	Medium	19/04/2024	24/04/2024	Done

Table 3. **Project schedule.** Overview of the timeline for the Sign Language Translation System thesis

## Chapter 5

# MATERIALS AND METHODS

The Materials and Methods section in this thesis acts as a comprehensive guide outlining the research methodology implemented to achieve the project’s objectives. It delineates the data preprocessing steps, architectural designs of three approaches, implementation setup, evaluation metrics, and techniques utilized throughout the study. This section aims to enhance transparency and reproducibility in the research process by providing detailed insights into the experimental setup, data collection procedures, and analytical methods. By meticulously documenting these aspects, the section ensures the robustness and rigor of the research methodology. Additionally, it facilitates future replication of the study, thereby contributing to the advancement of knowledge in the field.

## 5.1 Data

How2Sign [88] is considered one of the most extensive datasets available in the realm of Sign Language, with a particular focus on American Sign Language (ASL). This multi-modal and multi-view dataset comprises over 80 hours of instructional videos covering a diverse range of topics. Derived from the renowned How2 dataset [89], a publicly available multi-modal dataset for vision, speech and natural language understanding, How2Sign offers ASL translations of more than 2500 instructional videos, each meticulously aligned with English transcripts. Enriched with annotations including category labels, English transcripts, gloss annotation, speech, depth data, and automatically extracted 2D body poses with OpenPose [43], it provides a comprehensive resource for research and development in ASL processing.

This study primarily utilized videos and their subtitles for both training and evaluating the models. Video clips were segmented based on timestamps provided by How2Sign

	<b>train</b>	<b>val</b>	<b>test</b>
Original	31,128	1,741	2,322
Final	30,942	1,734	2,341

Table 4. **How2Sign Dataset Overview.** Statistics of the original and final versions of the How2Sign dataset.

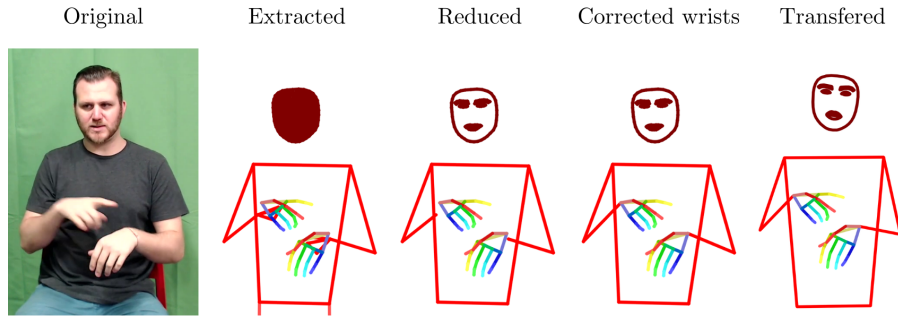


Figure 1. **Data preprocessing pipeline.** Left to right: original image, original pose extracted by Mediapipe Holistic, pose removed unnecessary landmarks, pose aligned hand wrists with body wrists, and final pose after transferring appearance.

(realigned version), combined with subtitles to create a comprehensive set of sentence-level data. As per the original paper, there are **31,128**, **1,741**, and **2,322** video-subtitle pairs in the training, validation, and test sets, respectively. However, upon initial examination, a small portion of the manually realigned subtitles were found to be invalid (i.e. exhibiting no temporal overlap with the video or falling outside the video duration). Additionally, the number of data in the manually realigned file differs from that stated in the original paper. After filtering and processing, the final splits utilized in this project were: **30,942** training videos, **1,734** validation videos, and **2,341** test videos. Table 4 presents the quantity of data in both the training, validation, and test sets before and after processing. Notably, although the signing videos encompass multi-view perspectives, this project utilized only the frontal view set due to resource limitations and project implementation timelines.

### 5.1.1 Data Preprocessing

In this study, the collected ground truth (GT) sign pose sequences were employed by extracting estimated pose landmarks from each video utilizing Mediapipe Holistic [90]. This state-of-the-art pose estimation framework accurately determines the 3D coordinates of various keypoints on the human body, encompassing the face, hands, and body. Each frame yielded 543 landmarks (33 pose landmarks, 468 face landmarks, and 21 hand landmarks per hand) extracted for a single person. Subsequently, the extracted 3D landmarks underwent other preprocessing steps, facilitated by the pose-format library [91], to generate the final training data. The complete data preprocessing pipeline is outlined in Figure 1. Other processing steps encompass:

1. **Mask unnecessary landmarks.** Due to the limited contribution of legs, the overlap between body hands and hands, and the abundance of facial landmarks (totaling 468 points), there was a risk of diminishing performance. To address this, the project implemented a filtering process to exclude landmarks associated with legs and body hands, while retaining only the crucial facial landmarks necessary for conveying emotions and mouth shape. Following this reduction, each individual was represented by a total of 178 remaining landmarks.
2. **Align hand wrists with body wrists.** Following the removal process, a gap emerged between hand wrists and body wrists. To address this issue, the project

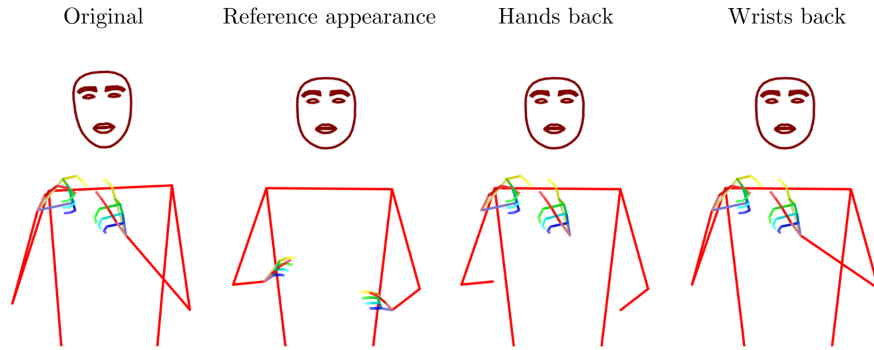


Figure 2. **Pose anonymization procedure.** Left to right: first frame, the reference appearance, pose brought back the hands, pose brought back the wrists.

resolves it by substituting body wrists with the corresponding hand wrists.

3. **Pose Normalization.** In this study, the pose keypoints are normalized using the pose shoulders to ensure uniform scale across all poses. The center of a pose is defined as the neck, determined by computing the average midpoint between the shoulders across all frames. Subsequently, all keypoints are translated, shifting the center to  $(0, 0)$ , and the pose is scaled so that the average distance between the shoulders equals 1.
4. **Pose Anonymization.** To remove identifying information from sign language poses and ensure data consistency for improved training efficiency, a pose anonymization process is implemented. This process involves assuming that the first frame solely depicts the person’s appearance, which is then removed from subsequent frames. Subsequently, a reference appearance is added. This study employed the reference appearance, determined as the mean of sign language poses calculated by [92]. Figure 2 illustrates the process of anonymization.

## 5.2 Methods

In this section, the architecture of three approaches will be elucidated. Given a spoken language sentence  $X = (x_1, \dots, x_T)$  with  $T$  words, the models generate a sequence of signs with  $\hat{Y} = (\hat{y}_1, \dots, \hat{y}_U)$  to closely resemble the ground truth  $Y = (y_1, \dots, y_U)$  with  $U$  time steps. Each sign pose frame,  $y_u$ , is represented as a continuous vector comprising the 3D joint positions of the signer.

### 5.2.1 Regressive Training with Progressive Transformers

The Autoregressive approach is embodied in the Progressive Transformers [12], chosen as the pioneer Sign Language Production (SLP) model capable of translating text into continuous 3D sign pose sequences in an end-to-end fashion. Developed in 2020 by Ben Saunders and colleagues [12], it marks a significant advancement in the field.

Due to the representation of sign language with continuous joint positions [14, 93], traditional symbolic Neural Machine Translation (NMT) architectures like Transformers cannot be directly applied without adaptation. This necessity led to the development



of a novel architecture: the progressive transformer-based architecture. It facilitates translation from symbolic input to a continuous output representation, as illustrated in Figure 3. To facilitate sequence length prediction of the continuous output, the authors introduce counter decoding, allowing the model to monitor sequence generation progress, hence the name Progressive Transformers. This approach also empowers timing control during inference, resulting in stable sign pose outputs, even in scenarios with no predefined vocabulary. In the following part of this section, a detailed description of the Progressive Transformers architecture is provided.

### Progressive Transformers

Progressive Transformers perform translation from symbolic text domains to continuous sign pose sequences, depicting the motion of a signer while producing a sentence in sign language. The model’s task is to generate skeleton pose outputs capable of both accurately translating the provided input sequence and realistically representing a sign pose sequence. First, it embeds the source tokens,  $x_t$ , through a linear symbolic embedding layer and joint values,  $y_u$ , through a linear continuous embedding layer, allowing similar content to be closely represented in the dense space. The symbolic and continuous embedding, involving weights  $W$  and bias  $b$ , can be expressed as:

$$\begin{aligned} w_t &= W^x \cdot x_t + b^x \\ j_u &= W^y \cdot y_u + b^y \end{aligned} \tag{1}$$

where  $w_t$  is the vector representations of the source tokens  $x^t$ ,  $y_u$  is the embedded 3D joint coordinates of each frame  $y_u$ .

Transformer networks lack inherent awareness of word order, as they receive all source tokens simultaneously, devoid of positional cues. To address this and introduce temporal ordering, a temporal embedding layer is applied after each input embedding. For the symbolic transformer, positional encoding [94] is implemented as:

$$\hat{w}_t = w_t + \text{PositionalEncoding}(t) \tag{2}$$

where *PositionalEncoding* is a predetermined sinusoidal function that depends on the relative sequence position, denoted by  $t$ .

With the sign poses, once they are embedded, a counter embedding layer is applied to them as temporal embedding (referred to as CE in Figure 3). The counter,  $c$ , ranges between 0 and 1, denoting the frame position relative to the total sequence length. The joint embeddings,  $j_u$ , are concatenated with their corresponding counter value,  $c_u$ , expressed as:

$$\hat{j}_u = [j_u, \text{CounterEmbedding}(u)] \tag{3}$$

where *CounterEmbedding* is a linear projection of the counter value for frame  $u$ .

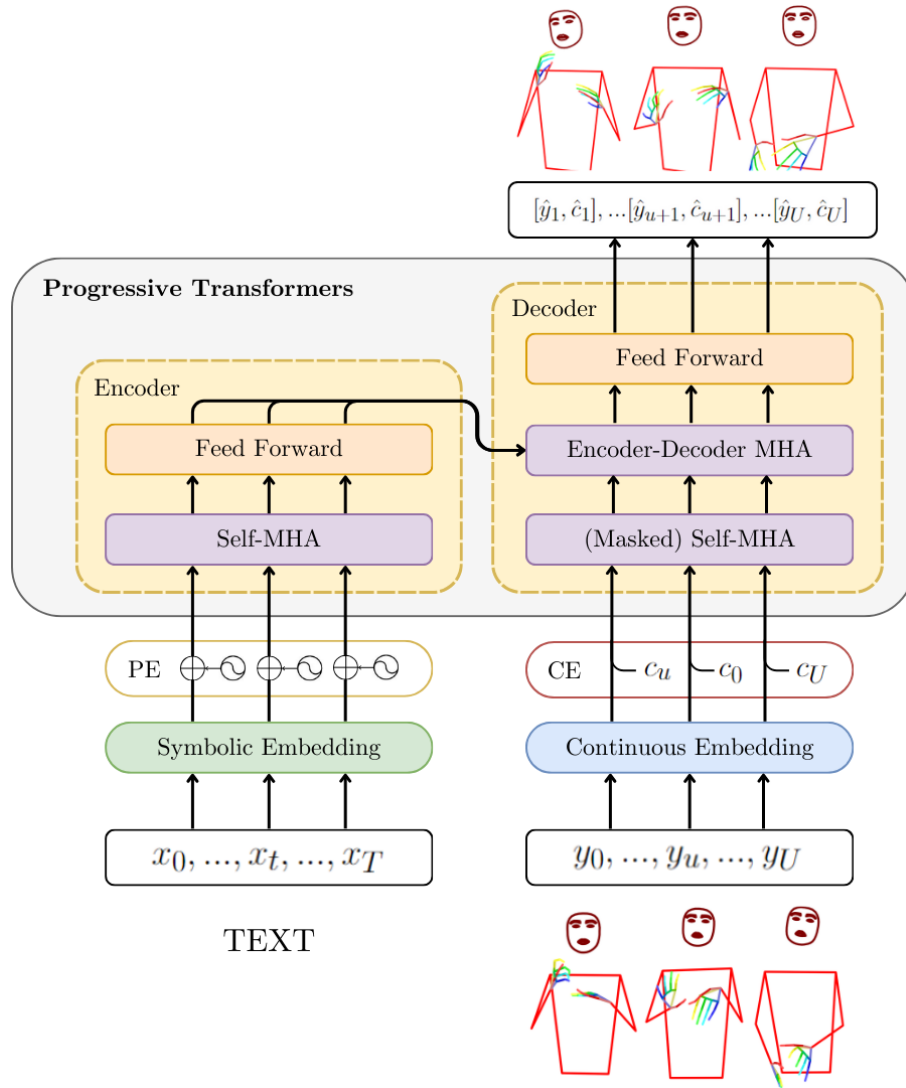


Figure 3. **Architecture details of Progressive Transformers.** (PE: Positional Encoding, CE: Counter Embedding, MHA: Multi-Head Attention).

During each time-step, counter values are forecasted alongside the skeleton pose, as illustrated in Figure 4. The sequence generation process halts once the counter attains a value of 1. This approach, termed Counter Decoding, allows for monitoring the progress of sequence generation and offers a means to anticipate the sequence's conclusion without relying on a tokenized vocabulary.

The counter furnishes the model with insights into the duration and speed of each sign pose sequence, influencing the timing of signs. During inference, the sequence generation process is guided by substituting the predicted counter value,  $\hat{c}$ , with the ground truth timing information,  $\hat{c}^*$ , thereby ensuring the production of a stable output sequence.

The Progressive Transformer adopts an encoder-decoder architecture. Within the symbolic encoder ( $E_S$ ), there exists a stack of  $L$  identical layers, each comprising 2 sub-layers. Commencing with the temporally encoded source embeddings,  $\hat{w}_t$ , a Multi-Head Attention (MHA) mechanism initially generates a weighted contextual representation by conducting multiple projections of scaled dot-product attention. This process aims

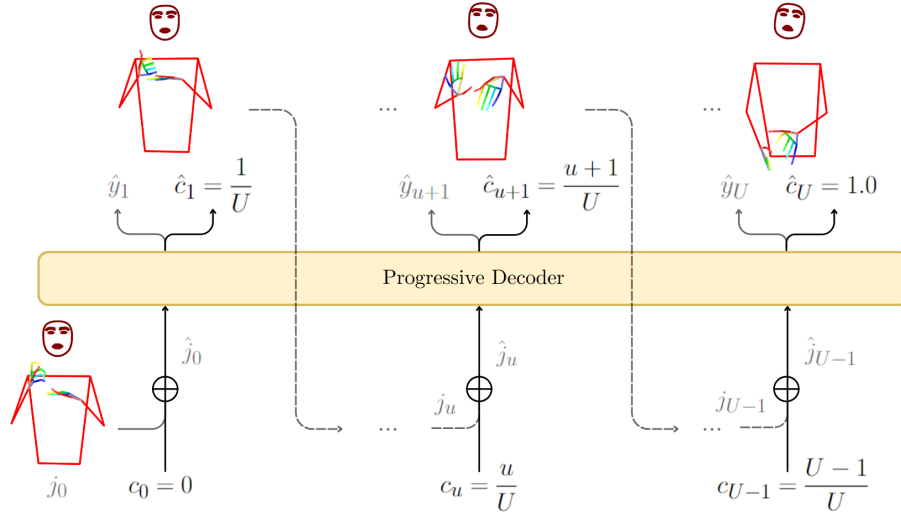


Figure 4. **An illustration of Counter Decoding.** Demonstrate the concurrent prediction of the sign pose,  $\hat{y}_u$ , and the counter,  $\hat{c}_u$ , where  $\hat{c}_u \in 0 : 1$ , and  $\hat{c}_u = 1.0$  denotes the end of the sequence.

to discern the relationship between each token of the sequence and its relevance at each time step within the entirety of the sequence. Formally, scaled dot-product attention yields a vector combination of values,  $V$ , weighted by relevant queries,  $Q$ , keys,  $K$ , and dimensionality,  $d_k$ :

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

The *MHA* module stacks parallel attention mechanisms in  $h$  different mappings of the same queries, keys, and values, each equipped with distinct learned parameters. This design facilitates the generation of diverse representations of the input, capturing complementary information across different sub-spaces. The outputs of each head are subsequently concatenated and projected forward through a final linear layer, expressed as:

$$\text{MHA}(Q, K, V) = [\text{head}_1, \dots, \text{head}_h] \cdot W^O \quad (5)$$

where  $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$  and  $W^O, W_i^Q, W_i^K$  and  $W_i^V$  are weights related to each input variable.

The outputs from the *MHA* mechanism are then passed into the second sub-layer, which involves a non-linear feed-forward projection. To facilitate training, a residual connection [95] and subsequent layer normalization [96] are applied around each of these sub-layers. The final output of the symbolic encoder can be expressed as:

$$r_t = E_S(\hat{w}_t | \hat{w}_{1:T}) \quad (6)$$

where  $E_S$  is the symbolic encoder and  $r_t$  is the encoded source representation.

The progressive decoder ( $D_P$ ) operates as an auto-regressive model, generating a sign pose frame and its corresponding counter value at each time-step, as explained previously. The progressive decoder generates continuous sequences characterized by sparse representation within a vast continuous subspace. The joint embeddings,  $\hat{j}_u$ , concatenated with the counter, serve as the target input, encapsulating the sign information for each frame.

Initially, a self-attention Multi-Head Attention ( $MHA$ ) sub-layer is applied, incorporating target masking to prevent attending to future positions. Subsequently, another  $MHA$  mechanism is utilized to transform the symbolic representations from the encoder into the continuous domain of the decoder, facilitating the learning of crucial alignments between spoken and sign languages. Following this, a final feed-forward sub-layer is employed, with each sub-layer featuring a residual connection and layer normalization, as previously described. Unlike typical models, no softmax layer is utilized, as the skeleton joint coordinates can be directly regressed without the need for stochastic prediction. The progressive decoder output is represented as:

$$[\hat{y}_{u+1}, \hat{c}_{u+1}] = D_P(\hat{j}_u | \hat{j}_{1:u-1}, r_{1:T}) \quad (7)$$

where  $\hat{y}_{u+1}$  denotes the 3D joint positions representing the produced sign pose of frame  $u + 1$ , while  $\hat{c}_{u+1}$  represents the respective counter value. The decoder is trained to generate one frame at a time until the predicted counter value reaches 1, signaling the end of the sequence. Upon completing the full sign pose sequence, the model undergoes end-to-end training utilizing Mean Squared Error ( $MSE$ ) loss, calculated between the predicted sequence,  $\hat{y}_{1:U}$ , and the ground truth,  $y_{1:U}^*$ :

$$\mathcal{L}_{MSE} = \frac{1}{U} \sum_{i=1}^u (y_{1:U}^* - \hat{y}_{1:U})^2 \quad (8)$$

The outputs of the progressive transformer,  $\hat{y}_{1:U}$ , represent the 3D skeleton joint positions of each frame within a produced sign sequence.

### 5.2.2 Adversarial Training with Progressive Transformers

Expanding on the foundation laid by the Progressive Transformer architecture [12], Saunders and colleagues introduced Adversarial Training for Multi-Channel SLP [16]. This innovative approach incorporates a conditional adversarial discriminator alongside the regression loss, aiming to mitigate the issues of regression to the mean and prediction drift inherent in the original architecture.

Sign language encompasses subtle and precise movements of both manual and non-manual components. However, existing SLP models often suffer from regression to the mean, leading to under-articulated output characterized by average hand shapes due to the variability of joint positions. To tackle this challenge, the authors propose an adversarial training mechanism for SLP. This method utilizes the Progressive Transformer architecture described earlier as a Generator, denoted as  $G$ , to generate sign pose sequences from input text. To ensure the production of realistic and expressive sign language, a

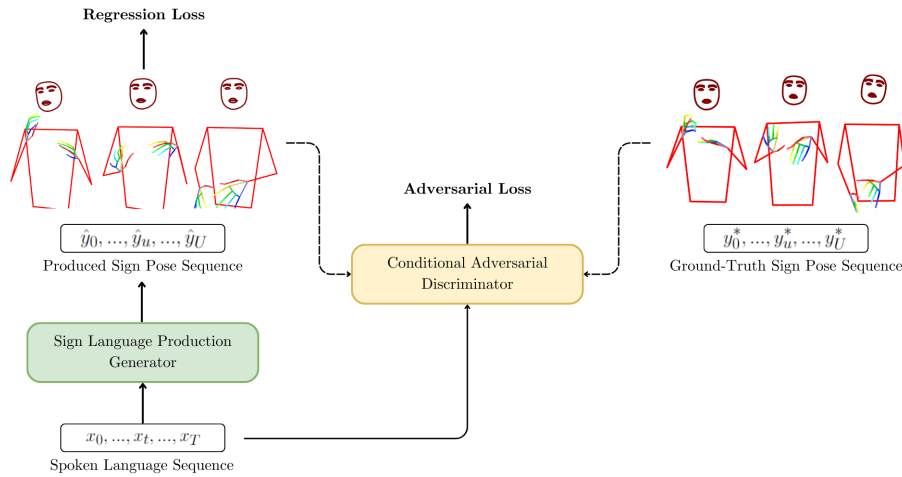


Figure 5. **An overview of Adversarial Multi-Channel SLP.** Featuring a Conditional Adversarial Discriminator assessing the realism of Sign Pose Sequences generated by an SLP Generator.

conditional adversarial Discriminator,  $D$ , is introduced. This discriminator learns to distinguish between real and generated sign pose sequences conditioned on the input spoken language. Both models are then co-trained in an adversarial manner, resulting in mutually improved performance. An overview of this approach is shown in Figure 5. Formally, the adversarial training framework for SLP can be defined as a minimax game, with  $G$  striving to minimize the following equation while  $D$  aims to maximize it:

$$\min_G \max_D \mathcal{L}_{GAN}(G, D) = \mathbb{E}[\log D(Y^*|X)] + \mathbb{E}[\log(1 - D(G(X))|X)] \quad (9)$$

where  $Y^* = y_{1:U}^*$  is the ground truth sign pose sequence,  $G(X)$  equates to the produced sign pose sequence,  $\hat{Y} = \hat{y}_{1:U}$ , and  $X$  is the source spoken language.

In the remainder of this section, a comprehensive overview of both the Generator and Discriminator will be presented.

## The Generator

The Generator, denoted as  $G$ , is trained to generate sign pose sequences based on a given source spoken language sequence, incorporating the progressive transformer within a GAN framework. Unlike the conventional GAN setup, this implementation demands that sequence generation be conditioned on a particular source input. Consequently, the traditional noise input [97] is omitted, and a sign pose sequence is generated based on the source sequence, drawing inspiration from conditional GANs [98].

During training,  $G$  is trained using a combination of loss functions, comprising the regression loss  $\mathcal{L}_{Reg}$  - the Mean Squared Error (MSE) loss functions of the Progressive Transformer (Equation 8), and the adversarial loss  $\mathcal{L}_{GAN}^G$  (Equation 9). The total loss function is formulated as:

$$\mathcal{L}^G = \lambda_{Reg} \mathcal{L}_{Reg}(G) + \lambda_{GAN} \mathcal{L}_{GAN}^G(G, D) \quad (10)$$

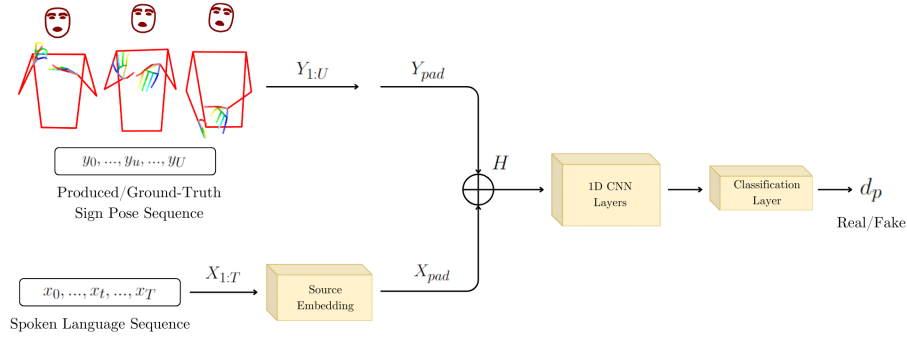


Figure 6. **Details of our Conditional Adversarial Discriminator architecture.** The sign pose sequence  $Y_{1:U}$  is concatenated with the source spoken language sequence  $X_{1:T}$  and mapped to a single scalar,  $d_p$ .

where  $\mathcal{L}_{GAN}^G$  represents the latter component of Equation 9 and  $\lambda_{Reg}$ ,  $\lambda_{GAN}$  determine the relative importance of each loss function during training. The regression loss offers specific guidance on generating the given input, while the adversarial loss guarantees realistic signer motion. Together, these losses collaborate to achieve both accurate and expressive sign production.

## The Discriminator

A conditional adversarial Discriminator, denoted as  $D$ , is employed to distinguish between generated sign sequences,  $\hat{Y}$ , and ground-truth sign sequences,  $Y^*$ , conditioned on the source spoken language sequence,  $X$ . The objective of  $D$  is to assess the realism of sign production, thereby guiding  $G$  towards producing expressive and articulate outputs. Additionally, conditioning on the source sequence enables  $D$  to simultaneously evaluate the translation accuracy of the source-target sequence pair,  $(X, Y)$ . An overview of the discriminator architecture is illustrated in Figure 6.

For every pair of source-target sequences,  $(X, Y)$ , whether they are generated or real sign pose sequences, the goal of the discriminator is to generate a single scalar,  $d_p \in (0, 1)$ , which represents the probability that the sign pose sequence originates from the ground-truth data,  $Y^*$ :

$$d_p = P(Y = Y^* | X, Y) \in (0, 1) \quad (11)$$

To accommodate the variable frame lengths of the sign sequences, padding is applied to standardize them to a fixed length, denoted as  $U_{max}$ , representing the maximum frame length observed in the target sequences within the data:

$$Y_{pad} = [y_{1:U}, \emptyset_{T:T_{max}}] \quad (12)$$

where  $Y_{pad}$  is the sign pose sequence padded with zero vectors  $\emptyset$  to facilitate convolutions on a tensor of fixed size. To condition the discriminator on the source spoken language, we initially embed the source tokens using a linear embedding layer. Dealing with variable sequence lengths, these embeddings are also padded to a fixed length, denoted as  $T_{max}$ , which represents the maximum source sequence length:

$$X_{pad} = [W^X \cdot X_{1:T} + b^X, \emptyset_{T:T_{max}}] \quad (13)$$

where  $W^X$  and  $b^X$  represent the weight and bias of the source embedding, respectively, and  $\emptyset$  denotes zero padding. As depicted in the center of Figure 2, the source representation is subsequently concatenated with the padded sign pose sequence to form the conditioned features, denoted as  $H$ :

$$H = [Y_{pad}, X_{pad}] \quad (14)$$

To assess the realism of the sign pose sequence, the discriminator extracts meaningful representations through multiple layers of 1D CNNs. These convolutional filters traverse the sign pose sequence, analyzing the local context to gauge the temporal coherence of the signing motion. This approach proves more effective than employing a frame-level discriminator, as it evaluates the consistency of hand shapes over a larger temporal window rather than focusing on individual frames. Leaky ReLU activation [99] is applied after each layer to foster healthy gradients during training. Finally, a feed-forward linear layer and sigmoid activation project the combined features to a single scalar,  $d_p$ , representing the probability that the sign pose sequence is real.

The discriminator is trained by maximizing the likelihood of generating  $d_p = 1$  for real sign sequences and  $d_p = 0$  for generated sequences. This objective can be formalized as maximizing Equation 9, leading to the loss function  $\mathcal{L}^D = \mathcal{L}_{GAN}^D(G, D)$ .

### 5.2.3 Non-Autoregressive Transformers with Gaussian Space

Previous methods (5.2.1 and 5.2.2 approaches) focus on maximizing the conditional probability  $P(Y|X)$ . However, due to the curse of dimensionality [100] and the significant disparity in length between  $X$  and  $Y$ , an autoregressive approach proves ineffective.

To address these challenges from a foundational standpoint, a novel SLP model, Non-Autoregressive Sign Language Production with Gaussian space (NSLP-G) [17], is proposed. This model takes a distinct approach from existing SLP models by employing two distinct phases: constructing a pose generator and mapping from a source sentence to target sign pose distributions. In Phase I, Variational Autoencoder (VAE) is utilized for self-supervised learning on the sign poses. Following this learning phase, the decoder is capable of generating a sign pose in Gaussian space, thus serving as the Gaussian Pose Generator (GPG). In Phase II, a non-autoregressive Transformer acts as the Gaussian Seeker (GS), translating the source sentence into the target sign pose distributions based on the GPG. The key innovation of this model lies in furnishing the decoder with positional encoding and outputting the entire sign pose sequence at once.

In this method, direct regression is avoided, and instead, words  $X$  are mapped to sign pose distributions  $Z$ , where  $Z = (z_1, z_2, \dots, z_t)$  generates sign pose  $Y$  using a generator  $g(\cdot)$ . This formulation can be expressed as:

$$P(Z|X), \quad g(Y|Z), z_i \in N(0, 1) \quad (15)$$

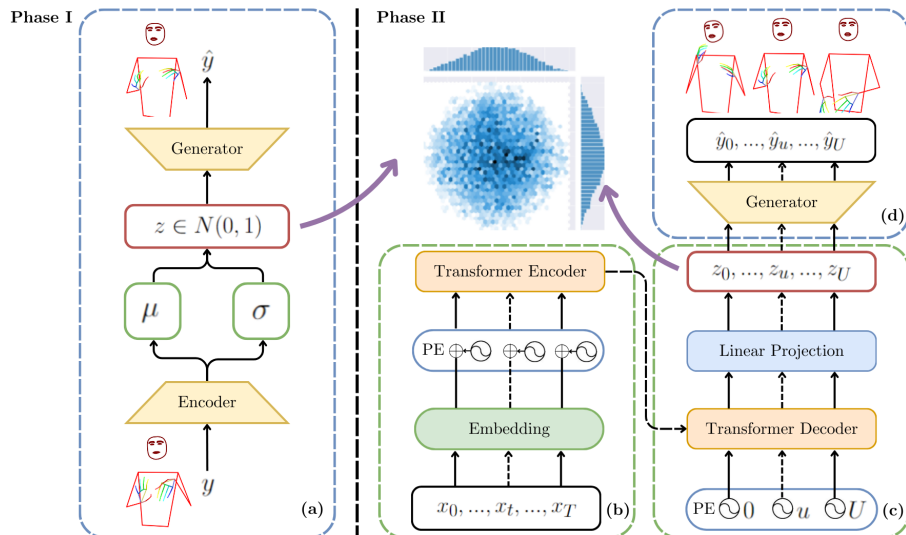


Figure 7. **An overview of the NSLP-G model.** Comprising two phases for generating sign poses based on a given source sentence. Phase I: VAE is utilized to serve as Gaussian Pose Generator (GPG). The encoder produces  $\mu$  and  $\sigma$ , and then employs a reparameterization trick to sample  $z$  following  $N(0, 1)$ . The decoder reconstructs  $\hat{y}$  using  $z$ . Phase II: To generate a sequence of  $z$ , a Transformer-equipped GPG is employed to produce a sequence of  $z$ . To achieve non-autoregressiveness, autoregressive connections are eliminated from the decoder, and only positional encodings (PE) are utilized as inputs.

To maximize  $P(Z|X)$  and  $g(Y|Z)$ , the authors employ Transformer and Variational Autoencoder (VAE), respectively. The specifics of each method are covered in the subsequent section.

### Gaussian Pose Generator with VAE

In Phase I, depicted in Figure 7(a), this method utilized VAE, a widely adopted technique for generative tasks [101, 102, 103, 104, 105], to acquire GPG. VAE is trained to generate a sign pose  $\hat{y}$  that closely resembles the ground truth sign pose  $y$ . It comprises a simple architecture with a sign pose encoder  $enc_{sp}$  and decoder  $dec_{sp}$  akin to Autoencoder (AE) [106]. The encoder processes a sign pose  $y$  and encodes it into latent space, while the decoder reconstructs a sign pose from the latent space  $z_{sp}$ . The encoder and decoder can be represented as follows:

$$enc_{sp}(y) = q_{sp}(z_{sp}|y), \quad dec_{sp}(z_{sp}) = p_{sp}(y|Z_{sp}) \quad (16)$$

where  $q_{sp}(z_{sp}|s)$  and  $p_{sp}(s|z_{sp})$  are the posterior distributions for the encoder and decoder, respectively.

VAE employs a reparameterization trick to sample the latent vector  $z_{sp}$  from the output of the encoder, allowing the sign pose  $y$  to be projected into Gaussian space. This reparameterization trick can be expressed as follows:



$$z_{sp} = \mu_{sp} + \sigma_{sp} \odot \varepsilon, \quad \text{where } \varepsilon \in N(0, 1) \quad (17)$$

where  $\mu_{sp}$  and  $\sigma_{sp}$  represent the mean and variance of the sign pose distribution, respectively;  $\varepsilon$  denotes an auxiliary independent random variable; and  $\odot$  signifies element-wise multiplication.

The loss function of VAE can be expressed as:

$$\mathcal{L}_{VAE}(y) = -\mathbb{E}_{z_{sp} \sim q_{sp}(z_{sp}|y)}[\log p_{sp}(y|z_{sp})] + \beta KL(q_{sp}(z_{sp}|y)||p_{sp}(z_{sp})) \quad (18)$$

where  $p(z_{sp}) = N(0, 1)$  represents the prior distribution, and  $KL(\cdot||\cdot)$  denotes the Kullback-Leibler (KL) divergence. The first term enables the model to encode the sign pose  $y$  into the latent space  $z_{sp} \in N(0, 1)$  for reconstruction. The Mean Squared Error (MSE) loss was employed to guide the decoder to assume a Gaussian distribution. The second term encourages the posterior distribution  $q_{sp}(z_{sp}|y)$  to closely align with the prior distribution  $p_{sp}(z_{sp})$ . Additionally, the authors introduce a variable weight  $\beta$ , defined by KL cost annealing [101]. Following the learning process, the trained decoder  $dec_{sp}$  is designated as the GPG (see Figure 7 (d)).

### Gaussian Seeker with Non-Autoregressive Transformer

In Phase II, depicted in Figure 8, a Transformer is constructed in a non-autoregressive manner and utilized as the Gaussian Seeker.

The Encoder employs the same Transformer architecture, comprising a stack of  $N$  identical layers with Multi-Head Attention (MHA) and Feed-Forward layers. For more detailed information about the Encoder, please refer to the section 5.2.1

The Decoder operates in a non-autoregressive manner, where the autoregressive mask is removed. The  $P(Z|X)$  in Equation 15 can be represented as follows:

$$P_{NA}(Z|X) = \prod_{u=1}^U p_{gs}(z_u|x_{1:T}) \quad (19)$$

where  $Z$  and  $W$  are a target sequence of sign poses and a source sentence, respectively.

These distributions can be computed simultaneously during inference. However, as Equation 19 illustrates, there is no conditional probability to predict the length of the target distribution sequence  $Z$ . The model generates a fixed sequence of sign poses while utilizing a masked Mean Squared Error (MSE) loss, enabling the model to learn sign poses of varying lengths. With this loss calculation, the model converges to an idle state upon completion of inference.

In detail, the decoder utilizes positional encodings (PE) as a query, and the encoder's output serves as key and value inputs. Similar to the Encoder, the decoder consists of the same number of layers, each containing Multi-Head Attention (MHA) self-attention,

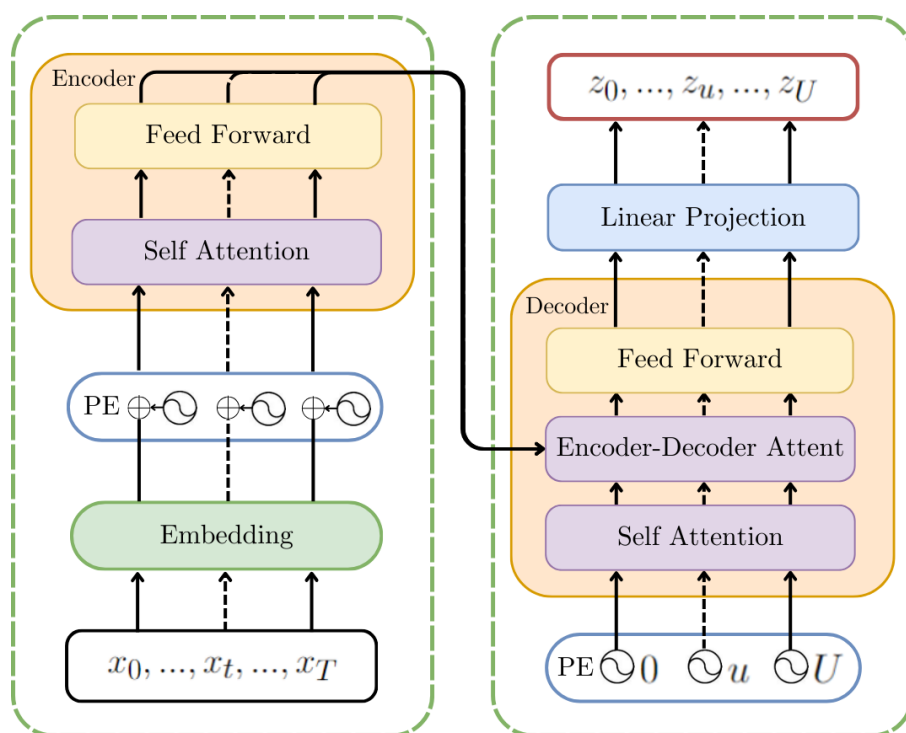


Figure 8. **A depiction of the Transformer-based Gaussian Seeker.** This module comprises a transformer encoder and a non-autoregressive decoder. The encoder processes the source sentence  $x_0, \dots, x_t, \dots, x_T$  consisting of  $U$  words, while the decoder takes positional encodings (PE) with a length of  $U$  as input to generate a sequence of latent vectors  $Z$  following a Gaussian distribution.

encoder-decoder attention, and feed-forward sub-layers. Ultimately, the decoder produces a target sequence  $Z$  via a linear projection layer.

### 5.3 Implementation Setup

This project setup closely follows the methodology outlined in the original works. It's worth noting that the original Progressive Transformer [12] employed several data augmentation techniques to address drift issues and significantly enhance the performance of SLP models. However, in this work, the model was trained and evaluated on the How2Sign [88] dataset without augmentation. This decision is motivated by the substantial volume of data available in the How2Sign dataset, which exceeds that of the RWTH-PHOENIX-Weather-2014T [7] dataset. The absence of augmentation allows for a direct comparison between models, serving the purpose of model comparison effectively.

### 5.4 Metrics

For evaluation, this work adopted the back-translation evaluation metric for SLP, as introduced by Saunders et al. [12]. This involved utilizing a pre-trained SLT model [107], which was trained on How2Sign dataset [88], to translate the generated sign pose sequences back into spoken language. This approach draws parallels to the use of inception score for generative models [108], which employs a pre-trained classifier. Additionally, BLEU and ROUGE scores were computed against the original input, with BLEU n-grams ranging from 1 to 4 provided for comprehensive evaluation.

## Chapter 6

# RESULTS

The Results chapter plays a crucial role in delivering the tangible outcomes of the project, achieved through the real-world training and evaluation processes of the 3 approaches outlined in the preceding chapter (Chapter 5). Within this chapter, the report will present experimental findings obtained from the research endeavors, with both quantitative and qualitative results detailed comprehensively and lucidly. Through these results, a comprehensive overview of the project's discoveries, valuable perspectives, and meaningful implications for further consideration and action will be provided in the subsequent chapter 7 Discussion.

### 6.1 Quantitative Results

The project conducted experiments on three approaches to the Text-to-Pose task, where a English sentence is inputted and the output consists of sequence of poses. The experiments were carried out following the implementation setup in the original works, using the back-translation evaluation metric (Section 5.4) to assess the performance of the three models on the validation and test sets of the How2Sign dataset.

As depicted in Table 5, the Adversarial Training regime notably enhances performance compared to the sole Progressive Transformers, which is trained solely with a Regression loss. Non-Autoregressive approach demonstrates the highest performance, outperforming the other two methods, which operate in an Autoregressive manner. The performance enhancement achieved with NSLP-G highlights a distinct gap.

### 6.2 Qualitative Results

In this qualitative results section, the report presents two cases with distinct purposes. The first case evaluates the three models using data with minimal representation of non-manual features (facial expressions, head movements, etc.), while the second case focuses primarily on evaluating data with strong non-manual features. Both cases utilize data from the test set of the How2Sign dataset.

In the first case, as illustrated in Figure 9, all three approaches predominantly succeed in translating text into the sign pose sequence. It is observed that the autoregressive

Models	VAL SET					TEST SET				
	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE
Progressive Transformers [12]	13.24	17.12	22.63	34.19	36.51	12.35	15.61	21.98	32.71	34.93
Adversarial Training with PT [16]	14.25	18.29	24.06	35.61	38.04	12.97	16.33	22.89	33.45	34.90
NSLP-G [17]	<b>15.83</b>	<b>20.15</b>	<b>26.14</b>	<b>37.34</b>	<b>40.05</b>	<b>15.71</b>	<b>18.84</b>	<b>26.93</b>	<b>37.62</b>	<b>39.52</b>

Table 5. **Performance comparison of three approaches to Text-to-Pose task.** Results indicate that the Adversarial Training regime significantly enhances performance compared to the Progressive Transformers trained solely with a Regression loss. The Non-Autoregressive approach, particularly NSLP-G, demonstrates the highest performance, showcasing a distinct advantage over the other methods, which operate in an Autoregressive manner.

approaches of the first two methods exhibit how the preceding pose influences the subsequent pose. For Progressive Transformers (PT), noticeable deviations from the original pose begin from the fourth pose, resulting in a sequence with accumulating errors. Similarly, with the second approach, Adversarial Training with PT, deviations start from the seventh pose. Conversely, with the Non-Autoregressive approach, except for the fifth and sixth sign poses, the generated poses exhibit a certain degree of accuracy. This indicates that the non-autoregressive decoder effectively generates the next sign pose without carrying forward errors from previously generated sign poses.

In the second case, the results somewhat corroborate the observations from the first case. Regarding non-manual features, it can be observed that the models also partially reproduce them, with NSLP-G leading in closest resemblance to the original. However, due to limitations of pose type visualization, non-manual features are not as clearly depicted compared to other data modalities like RGB images, as depicted in Figure 10.

It’s worth noting that in both cases, the models produce relatively close results to the original videos. Nonetheless, discrepancies still exist in hand positions, finger orientations, and so on.

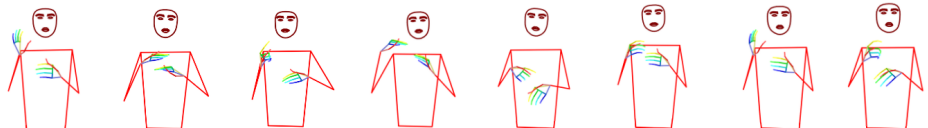
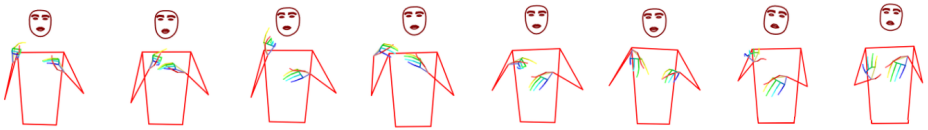
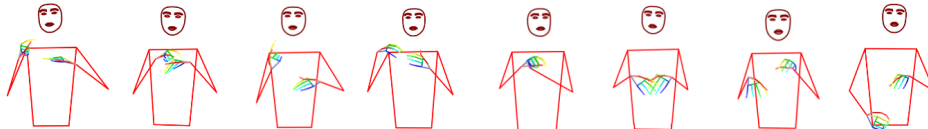

Progressive Transformers	
Adversarial Training with Progressive Transformers	
Non-Autoregressive with Gaussian Space	
Original Video	
Input text	Let the wrist do all the leading.

Figure 9. **Visualization of generated sign pose sequences in Case 1 evaluation.** Observations indicate varying levels of accuracy and deviation from the original poses across the models, highlighting the influence of autoregressive and non-autoregressive approaches on pose generation.

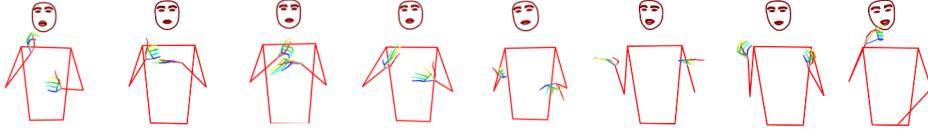
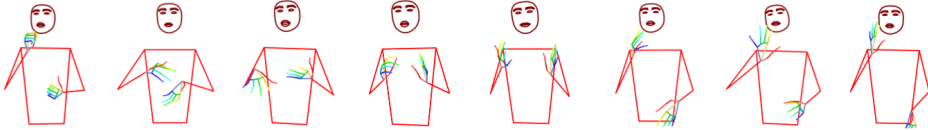
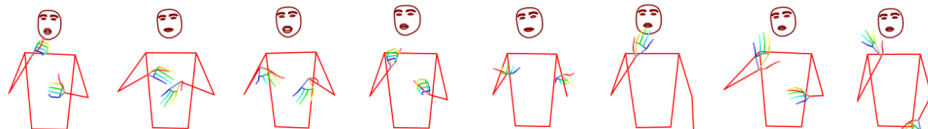

Progressive Transformers	
Adversarial Training with Progressive Transformers	
Non-Autoregressive with Gaussian Space	
Original Video	
Input text	The most expensive thing is probably going to be your food and drinks.

Figure 10. **Visualization of generated sign pose sequences in Case 2 evaluation.** Despite limitations in visualizing poses, certain non-manual features such as facial expressions and head movements are partially evident in the generated sequences. NSLP-G exhibits the closest resemblance to the original non-manual features among the evaluated models.

## Chapter 7

# DISCUSSIONS

The Discussion chapter serves as a platform for thorough analysis, interpretation, and exploration of the implications arising from the preceding chapters' results 6. This chapter delves into the nuances of the findings, uncovering both insights and limitations inherent in the research process. Moreover, it explores potential avenues for future research and practical applications, thereby fostering a more comprehensive understanding of the subject matter and enriching the broader discourse within the field. By presenting clear analyses, the project contributes significantly to the discussion surrounding sign language production, promoting deeper understanding and providing insights for future endeavors in this domain.

## 7.1 Interpretation and Implications

The project's achievement of higher results compared to those achieved with original papers is evident, which can be attributed to several factors, primarily the differences in the training dataset and the back-translation evaluation metric. It's worth noting that How2Sign is a significantly larger and more diverse dataset compared to the PHOENIX14T dataset used by other authors, models developed on this dataset also yield higher results.

Regarding Regression Training (Approach 1) vs Adversarial Training (Approach 2), the results indicate that the adversarial training regime improves performance over and regression training with Progressive Transformers architecture, aligning with the conclusions drawn in the original paper [16]. This demonstrates that incorporating a discriminator model in the second approach significantly enhances sign production comprehension. As the discriminator is conditioned upon the source text, the generator is prompted to accomplish both accurate translation and realistic production tasks simultaneously. Adversarial training also yields a close correspondence to the ground truth video, alongside accurate mouthings and head movements, with hand shapes becoming more expressive and meaningful.

In the comparison between Autoregressive (Approach 1, 2) and Non-Autoregressive (Approach 3), the results show that the non-autoregressive model (NSLP-G) outperforms the autoregressive models. The primary advantage of NSLP-G lies in its use of a well-constructed Gaussian space to produce sign poses in parallel, enabling the model to

generate the next sign pose without propagating errors from previously produced sign poses. In contrast, autoregressive models' output is significantly influenced by previous poses, resulting in more discrepancies. The autoregressive model can obtain excellent performance, while the non-autoregressive model brings fast decoding speed for inference. However, autoregressive models produce sequences of increased articulation, with smoother production. Nonetheless, post-processing may be required to achieve smooth pose transitions. Another noteworthy point is that NSLP-G can generate more dynamic and accurate sign poses, especially facial expressions, enhancing the realism of the sign poses.

## 7.2 Limitations and Future Works

While the models have shown promising results, there are several limitations that need to be addressed in future research. Firstly, the predicted hand shapes or movements may not always be entirely accurate (see in Figure 9 and 10) due to missing or incorrect keypoints in the processed data. Improving the quality of pose estimation is crucial to enhancing the model's ability to interpret poses correctly. Future work could focus on refining hand pose estimation techniques to mitigate this issue.

In some cases, the models sometimes generate movements correctly but not precisely in the right location due to local proximity. This inconsistency needs to be addressed to ensure the accuracy of the generated sign sequences.

With evaluation metric, the back-translation has limitations in measuring the performance of the generated sign poses. This is because it heavily relies on the performance of the Sign Language Translation (SLT) model [107], which may not always be stable. Developing a new and stable metric specifically tailored for Sign Language Production (SLP) models could overcome this challenge.

Furthermore, the scarcity of data in sign language remains a significant obstacle. Increasing the amount of data is essential for improving the performance of the models. Future efforts should focus on expanding sign language datasets to enhance model training and generalization.

Finally, the generated pose sequences produced in this project could serve as a foundation for future research. These sequences could be utilized to animate avatars [109, 110] or condition Generative Adversarial Networks (GANs) [111, 112], opening up opportunities for more advanced applications in sign language processing.

Addressing these limitations and exploring future research directions will contribute to the continued advancement of sign language processing technology, ultimately improving communication accessibility for the Deaf and hard-of-hearing communities.



## Chapter 8

# CONCLUSIONS

In conclusion, this thesis has explored and evaluated various approaches to Sign Language Production (SLP) with a focus on American Sign Language (ASL). Through the experimentation and analysis of three different methods on the How2Sign dataset, valuable insights have been gained into the strengths and limitations of each approach. The results demonstrate that the Non-Autoregressive Transformers with Gaussian Space approach proves to be highly effective, particularly with languages that exhibit high complexity, such as sign language. Building upon these findings, the project developed a real-time inference demo with the NSLP-G model.

The results of the study demonstrate the potential of SLP models in generating sign language sequences from textual inputs. Despite the promising performance exhibited by the models, several challenges and limitations have been identified, including inaccuracies in hand shape and movement prediction, as well as the instability of back-translation evaluation metrics for SLP. Moving forward, future research efforts could focus on addressing these challenges by improving hand pose estimation techniques, refining model architectures, and developing more reliable evaluation metrics. Additionally, the expansion of sign language datasets and the incorporation of 3D annotations could further enhance the performance and applicability of SLP models.

Overall, this thesis contributes to the advancement of SLP technology and underscores the importance of accessibility and inclusivity in communication for the Deaf and hard-of-hearing communities. By continuing to innovate and refine SLP methodologies, the field can progress towards a future where sign language communication is more accessible and universally embraced.

## Appendix A

# Demo

In this chapter, details about the demo application will be presented. It's important to note that the demo has been designed specifically to serve the purpose of the graduation thesis.

### A.1 UI Structure

The main UI will be divided into 2 parts:

1. The random sample panel.
  - 1A. A text from a random chosen sample.
  - 1B. A sign language video corresponding to the above mentioned text.
2. The translation panel.
  - 2A. An text input for users to enter their text.
  - 2B. A pose sign language translated from the above users' text.

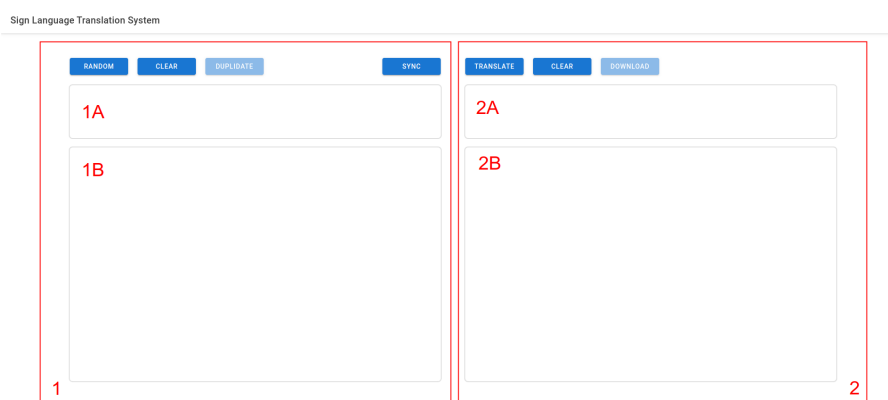


Figure 11. **The demo UI structure.** The UI includes 2 main parts.

## A.2 Use cases

### A.2.1 Get a random sample of text and its sign language video

To get a random sample, users can click on the Random button. A random text and its sign language video will be displayed on the right side panel.

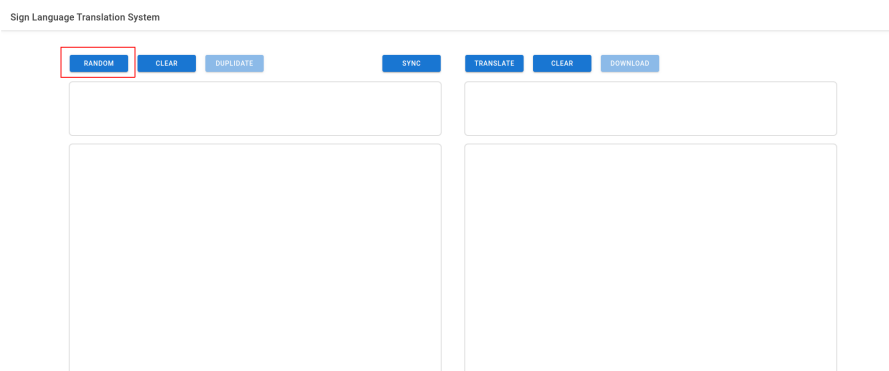


Figure 12. Use case with Random button. Users click on Random button.

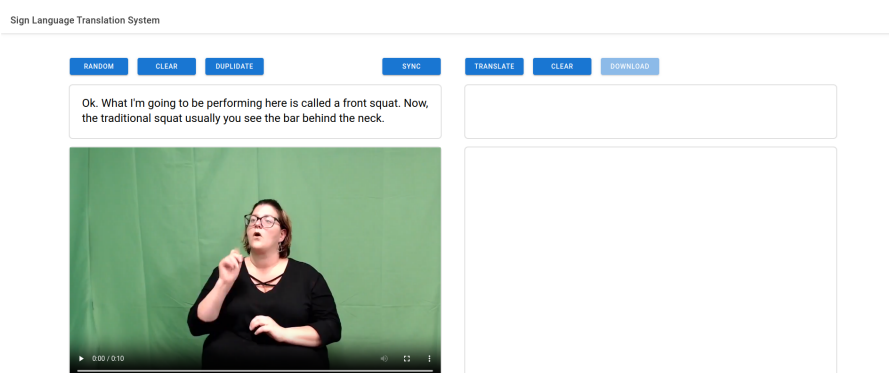


Figure 13. Use case with Random button. Result of user clicking the random button.

### A.2.2 Duplicate the text to translate panel input

To duplicate the text of random sample to translate panel input, users can click the Duplicate button.

### A.2.3 Translate the text to sign language pose

After users have filled out the input on the translate panel either by duplicating from random sample or by typing their own text, users can click the Translate button to translate from text to sign language pose video.

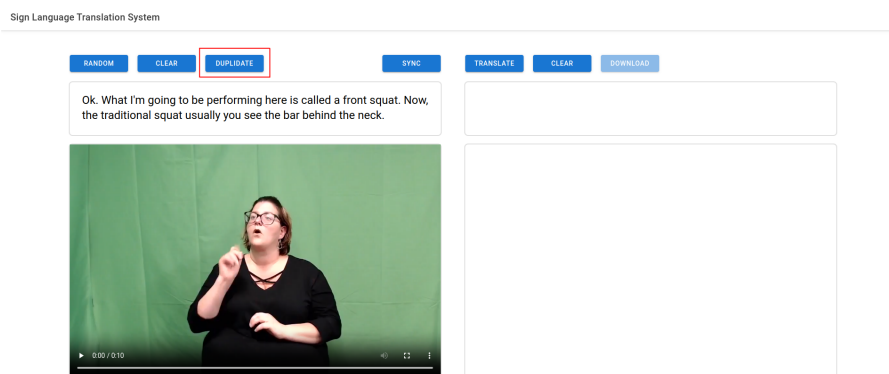


Figure 14. Use case with Duplicate button. Users click Duplicate button.

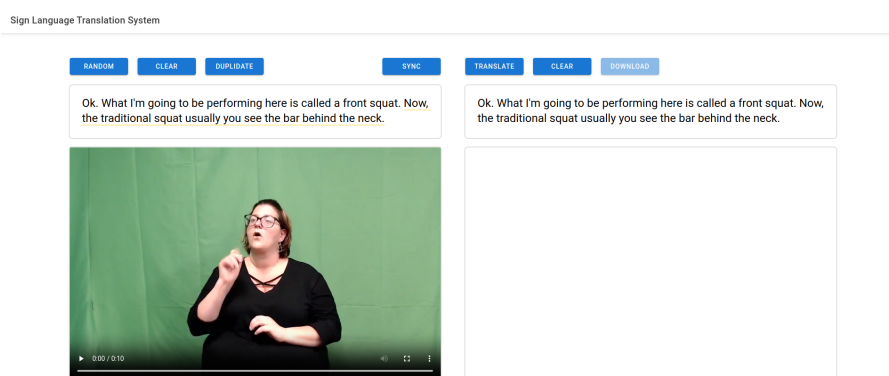


Figure 15. Use case with Duplicate button. Result of users clicking the Duplicate button.

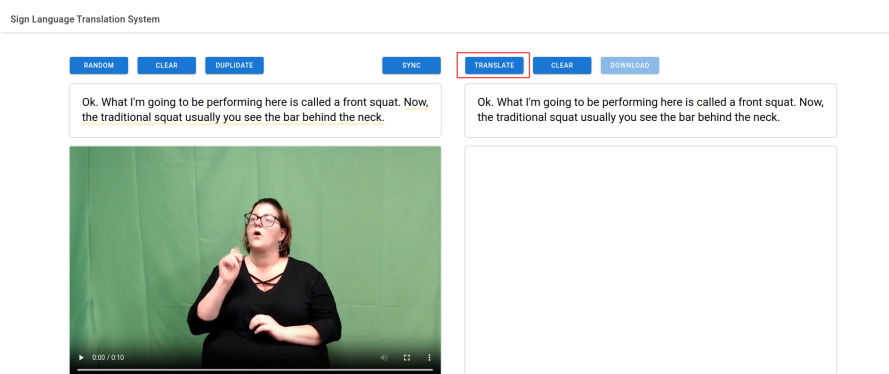


Figure 16. Use case with Translate button. Users click the Translate button.

#### A.2.4 Download the sign language pose video

After the pose video has been loaded successfully, users can click Download button to download the pose video

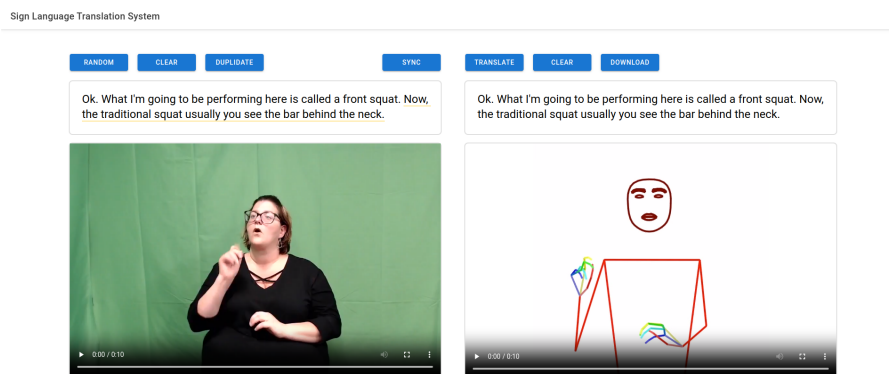


Figure 17. Use case with Translate button. Result of users clicking the Translate button.

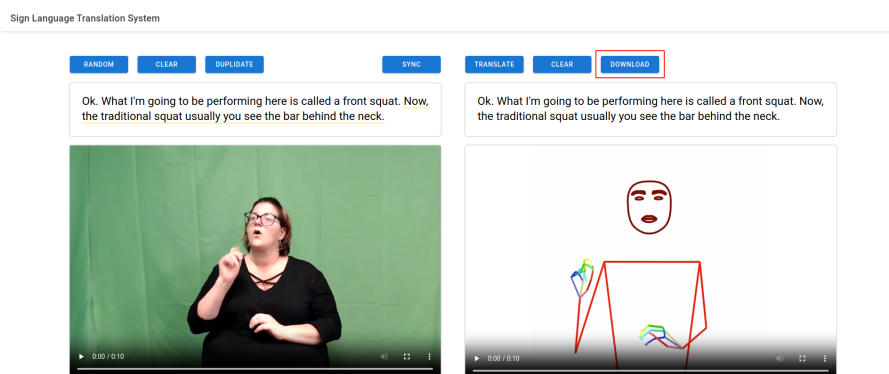


Figure 18. Use case with Download button. Users click the Download button.

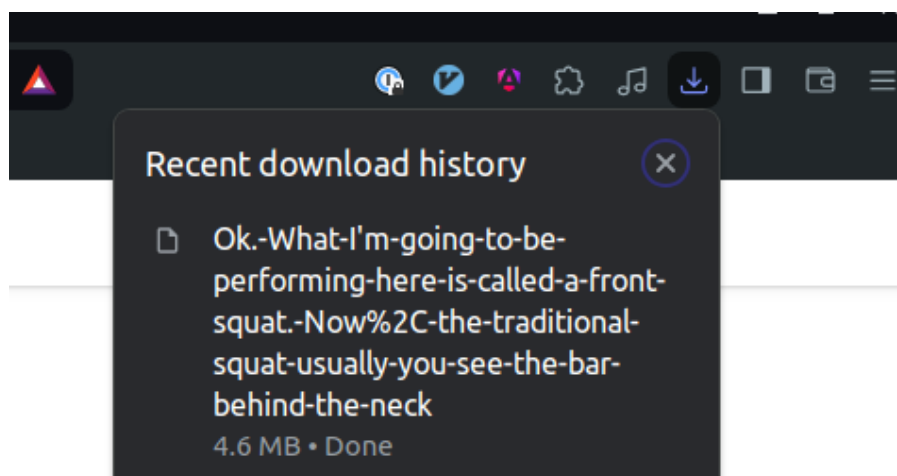


Figure 19. Use case with Download button. Result of uses clicking the Download button.

## REFERENCES

- [1] “Deafness and hearing loss.” <https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss>. Accessed: 15 April 2024.
- [2] “Sign language.” <https://education.nationalgeographic.org/resource/sign-language/>. Accessed: 15 April 2024.
- [3] W. Sandler and D. C. Lillo-Martin, *Sign language and linguistic universals*. Cambridge University Press, 2006.
- [4] C. A. Padden and T. L. Humphries, *Deaf in America*. Harvard University Press, 1988.
- [5] N. S. Glickman and W. C. Hall, *Language deprivation and deaf mental health*. Routledge, 2018.
- [6] S. Albanie, G. Varol, L. Momeni, T. Afouras, J. S. Chung, N. Fox, and A. Zisserman, “Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues,” in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 35–53, Springer, 2020.
- [7] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, “Neural sign language translation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7784–7793, 2018.
- [8] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, “Sign language transformers: Joint end-to-end sign language recognition and translation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10023–10033, 2020.
- [9] M. Parelli, K. Papadimitriou, G. Potamianos, G. Pavlakos, and P. Maragos, “Exploiting 3d hand pose estimation in deep learning-based sign language recognition from rgb videos,” in *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pp. 249–263, Springer, 2020.
- [10] R. Rastgoo, K. Kiani, and S. Escalera, “Hand sign language recognition using multi-view hand skeleton,” *Expert Systems with Applications*, vol. 150, p. 113336, 2020.
- [11] R. Rastgoo, K. Kiani, and S. Escalera, “Real-time isolated hand sign language recognition using deep networks and svd,” *Journal of Ambient Intelligence and*

- Humanized Computing*, vol. 13, no. 1, pp. 591–611, 2022.
- [12] B. Saunders, N. C. Camgoz, and R. Bowden, “Progressive transformers for end-to-end sign language production,” in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 687–705, Springer, 2020.
- [13] B. Saunders, N. C. Camgoz, and R. Bowden, “Mixed signals: Sign language production via a mixture of motion primitives,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1919–1929, 2021.
- [14] S. Stoll, N. C. Camgöz, S. Hadfield, and R. Bowden, “Sign language production using neural machine translation and generative adversarial networks,” in *Proceedings of the 29th British Machine Vision Conference (BMVC 2018)*, British Machine Vision Association, 2018.
- [15] S. Stoll, N. C. Camgoz, S. Hadfield, and R. Bowden, “Text2sign: towards sign language production using neural machine translation and generative adversarial networks,” *International Journal of Computer Vision*, vol. 128, no. 4, pp. 891–908, 2020.
- [16] B. Saunders, N. C. Camgoz, and R. Bowden, “Adversarial training for multi-channel sign language production,” *arXiv preprint arXiv:2008.12405*, 2020.
- [17] E. J. Hwang, J.-H. Kim, and J. C. Park, “Non-autoregressive sign language production with gaussian space,” in *BMVC*, vol. 1, p. 3, 2021.
- [18] S. K. Liddell and R. E. Johnson, “American sign language: The phonological base,” *Sign language studies*, vol. 64, no. 1, pp. 195–277, 1989.
- [19] R. E. Johnson and S. K. Liddell, “Toward a phonetic representation of signs: Sequentiality and contrast,” *Sign Language Studies*, vol. 11, no. 2, pp. 241–274, 2011.
- [20] D. Brentari, “Sign language phonology,” *The handbook of phonological theory*, pp. 691–721, 2011.
- [21] W. Sandler, “The phonological organization of sign languages,” *Language and linguistics compass*, vol. 6, no. 3, pp. 162–182, 2012.
- [22] U. Bellugi and S. Fischer, “A comparison of sign language and spoken language,” *Cognition*, vol. 1, no. 2-3, pp. 173–200, 1972.
- [23] S. K. Liddell, *Grammar, gesture, and meaning in American Sign Language*. Cambridge University Press, 2003.
- [24] T. Johnston and A. Schembri, *Australian Sign Language (Auslan): An introduction to sign language linguistics*. Cambridge University Press, 2007.
- [25] C. Rathmann and G. Mathur, “A featural approach to verb agreement in signed languages,” 2011.
- [26] A. Schembri, K. Cormier, and J. Fenlon, “Indicating verbs as typologically unique constructions: Reconsidering verb’agreement’in sign languages,” *Glossa*, vol. 3,

- no. 1, p. 89, 2018.
- [27] P. G. Dudis, “Body partitioning and real-space blends,” 2004.
- [28] S. K. Liddell and M. Metzger, “Gesture in sign language discourse,” *Journal of pragmatics*, vol. 30, no. 6, pp. 657–697, 1998.
- [29] L. de Beuzeville, “Pointing and verb modification: the expression of semantic roles in the auslan corpus,” in *Workshop Programme*, p. 13, 2008.
- [30] J. Fenlon, A. Schembri, and K. Cormier, “Modification of indicating verbs in british sign language: A corpus-based study,” *Language*, pp. 84–118, 2018.
- [31] T. ALLA, “The classifier system in american sign language,” *Noun classes and categorization*, p. 181, 1986.
- [32] S. Wilcox and S. Hafer, “Rethinking classifiers. emmorey, k.(ed.).(2003). perspectives on classifier constructions in sign languages. mahwah, nj: Lawrence erlbaum associates. 332 pages. hardcover. 69.95.,” 2004.
- [33] C. B. Roy, *Discourse in signed languages*. Gallaudet University Press, 2011.
- [34] K. Cormier, S. Smith, and Z. Sevcikova-Sehyr, “Rethinking constructed action,” *Sign Language & Linguistics*, vol. 18, no. 2, pp. 167–204, 2015.
- [35] R. Battison, “Lexical borrowing in american sign language.,” 1978.
- [36] S. Wilcox, “The phonetics of fingerspelling,” *The Phonetics of Fingerspelling*, pp. 1–114, 1992.
- [37] D. Brentari and C. A. Padden, “Native and foreign vocabulary in american sign language: A lexicon with multiple origins,” in *Foreign vocabulary in sign languages*, pp. 87–119, Psychology Press, 2001.
- [38] C. A. Padden, “The asl lexicon,” *Sign language & linguistics*, vol. 1, no. 1, pp. 39–60, 1998.
- [39] K. Montemurro and D. Brentari, “Emphatic fingerspelling as code-mixing in american sign language,” *Proceedings of the Linguistic Society of America*, vol. 3, pp. 61–1, 2018.
- [40] A. Isard, “Approaches to the anonymisation of sign language corpora,” in *Proceedings of the LREC2020 9th workshop on the representation and processing of sign languages: Sign language resources in the service of the language community, technological challenges and application perspectives*, pp. 95–100, 2020.
- [41] L. Pishchulin, A. Jain, M. Andriluka, T. Thormählen, and B. Schiele, “Articulated people detection and pose estimation: Reshaping the future,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3178–3185, IEEE, 2012.
- [42] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang, “Adversarial posenet: A structure-aware convolutional network for human pose estimation,” in *Proceedings*



- of the *IEEE international conference on computer vision*, pp. 1212–1221, 2017.
- [43] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7291–7299, 2017.
- [44] R. A. Güler, N. Neverova, and I. Kokkinos, “Densepose: Dense human pose estimation in the wild,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7297–7306, 2018.
- [45] M. Kato, “A study of notation and sign writing systems for the deaf,” *Intercultural Communication Studies*, vol. 17, no. 4, pp. 97–114, 2008.
- [46] S. Prillwitz and H. Zienert, “Hamburg notation system for sign language: Development of a sign writing with computer application,” in *Current trends in European Sign Language Research. Proceedings of the 3rd European Congress on Sign Language Research*, pp. 355–379, 1990.
- [47] W. C. Stokoe Jr, “Sign language structure: An outline of the visual communication systems of the american deaf,” *Journal of deaf studies and deaf education*, vol. 10, no. 1, pp. 3–37, 2005.
- [48] J. Kakumasu, “Urubu sign language,” *International journal of American linguistics*, vol. 34, no. 4, pp. 275–281, 1968.
- [49] B. Bergman, *Tecknad svenska:[Signed Swedish]*. LiberLäromedel/Utbildningsförl., 1977.
- [50] J. Mesch and L. Wallin, “Gloss annotations in the swedish sign language corpus,” *International Journal of Corpus Linguistics*, vol. 20, no. 1, pp. 102–120, 2015.
- [51] T. Johnston and L. De Beuzeville, “Auslan corpus annotation guidelines,” *Auslan Corpus*, 2016.
- [52] R. Konrad, T. Hanke, G. Langer, S. König, L. König, R. Nishio, and A. Regen, “Public dgs corpus: Annotation conventions,” *Technical report, Project Note AP03–2018-01, DGS-Korpus project*, 2018.
- [53] M. Müller, Z. Jiang, A. Moryossef, A. Rios, and S. Ebling, “Considerations for meaningful sign language machine translation based on glosses,” *arXiv preprint arXiv:2211.15464*, 2022.
- [54] L. Momeni, G. Varol, S. Albanie, T. Afouras, and A. Zisserman, “Watch, read and lookup: learning to spot signs from multiple supervisors,” in *Proceedings of the Asian Conference on Computer Vision*, 2020.
- [55] L. Momeni, H. Bull, K. Prajwal, S. Albanie, G. Varol, and A. Zisserman, “Automatic dense annotation of large-vocabulary sign language videos,” in *European Conference on Computer Vision*, pp. 671–690, Springer, 2022.
- [56] G. Varol, L. Momeni, S. Albanie, T. Afouras, and A. Zisserman, “Read and attend: Temporal localisation in sign language videos,” in *Proceedings of the IEEE/CVF*

- Conference on Computer Vision and Pattern Recognition*, pp. 16857–16866, 2021.
- [57] Y. Cheng, F. Wei, J. Bao, D. Chen, and W. Zhang, “Cico: Domain-aware sign language retrieval via cross-lingual contrastive learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19016–19026, 2023.
- [58] A. Duarte, S. Albanie, X. Giró-i Nieto, and G. Varol, “Sign language video retrieval with free-form textual queries,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14094–14104, 2022.
- [59] H. Hu, W. Zhou, and H. Li, “Hand-model-aware sign language recognition,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, pp. 1558–1566, 2021.
- [60] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu, “Skeleton aware multi-modal sign language recognition,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3413–3423, 2021.
- [61] H. R. V. Joze and O. Koller, “Ms-asl: A large-scale data set and benchmark for understanding american sign language,” *arXiv preprint arXiv:1812.01053*, 2018.
- [62] D. Li, C. Rodriguez, X. Yu, and H. Li, “Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1459–1469, 2020.
- [63] Y. Chen, F. Wei, X. Sun, Z. Wu, and S. Lin, “A simple multi-modality transfer learning baseline for sign language translation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5120–5130, 2022.
- [64] Y. Chen, R. Zuo, F. Wei, Y. Wu, S. Liu, and B. Mak, “Two-stream network for sign language recognition and translation,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 17043–17056, 2022.
- [65] K. L. Cheng, Z. Yang, Q. Chen, and Y.-W. Tai, “Fully convolutional networks for continuous sign language recognition,” in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pp. 697–714, Springer, 2020.
- [66] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, “Multi-channel transformers for multi-articulatory sign language translation,” in *Computer Vision—ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pp. 301–319, Springer, 2020.
- [67] K. Yin and J. Read, “Better sign language translation with stmc-transformer,” *arXiv preprint arXiv:2004.00588*, 2020.
- [68] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, “Multilingual denoising pre-training for neural machine translation,” *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020.

- [69] Q. Xiao, M. Qin, and Y. Yin, “Skeleton-based chinese sign language recognition and generation for bidirectional communication between deaf and hearing people,” *Neural networks*, vol. 125, pp. 41–55, 2020.
- [70] B. Saunders, N. C. Camgoz, and R. Bowden, “Everybody sign now: Translating spoken language to photo realistic sign language video,” *arXiv preprint arXiv:2011.09846*, 2020.
- [71] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, “Everybody dance now,” in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5933–5942, 2019.
- [72] B. Saunders, N. C. Camgoz, and R. Bowden, “Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5141–5151, 2022.
- [73] T. Hanke, L. König, S. Wagner, and S. Matthes, “Dgs corpus & dicta-sign: The hamburg studio setup,” in *sign-lang@ LREC 2010*, pp. 106–109, European Language Resources Association (ELRA), 2010.
- [74] W. Huang, W. Pan, Z. Zhao, and Q. Tian, “Towards fast and high-quality sign language production,” in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3172–3181, 2021.
- [75] D. Bragg, O. Koller, M. Bellard, L. Berke, P. Boudreault, A. Braffort, N. Caselli, M. Huenerfauth, H. Kacorri, T. Verhoef, *et al.*, “Sign language recognition, generation, and translation: An interdisciplinary perspective,” in *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 16–31, 2019.
- [76] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Q. Yuan, and A. Thangali, “The american sign language lexicon video dataset,” in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–8, IEEE, 2008.
- [77] U. Von Agris, M. Knorr, and K.-F. Kraiss, “The significance of facial features for automatic sign language recognition,” in *2008 8th IEEE international conference on automatic face & gesture recognition*, pp. 1–6, IEEE, 2008.
- [78] R. Wilbur and A. C. Kak, “Purdue rvl-slll american sign language database,” 2006.
- [79] M. Zahedi, D. Keysers, T. Deselaers, and H. Ney, “Combination of tangent distance and an image distortion model for appearance-based sign language recognition,” in *Pattern Recognition: 27th DAGM Symposium, Vienna, Austria, August 31-September 2, 2005. Proceedings 27*, pp. 401–408, Springer, 2005.
- [80] P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, and H. Ney, “Speech recognition techniques for a sign language recognition system,” *hand*, vol. 60, p. 80, 2007.
- [81] U. von Agris and K.-F. Kraiss, “Signum database: Video corpus for signer-independent continuous sign language recognition,” in *sign-lang@ LREC 2010*,

- pp. 243–246, European Language Resources Association (ELRA), 2010.
- [82] A. Schembri, J. Fenlon, R. Rentelis, S. Reynolds, and K. Cormier, “Building the british sign language corpus,” 2013.
- [83] A. Braffort, L. Bolot, E. Chételat-Pelé, A. Choisier, M. Delorme, M. Filhol, J. Segouat, C. Verrecchia, F. Badin, and N. Devos, “Sign language corpora for analysis, processing and evaluation.,” in *LREC*, 2010.
- [84] E. Efthimiou, S.-E. Fotinea, T. Hanke, J. Glauert, R. Bowden, A. Braffort, C. Collet, P. Maragos, and F. Goudenove, “Dicta-sign: sign language recognition, generation and modelling with application in deaf communication,” in *sign-lang@ LREC 2010*, pp. 80–83, European Language Resources Association (ELRA), 2010.
- [85] E. Efthimiou, S.-E. Fotinea, T. Hanke, J. Glauert, R. Bowden, A. Braffort, C. Collet, P. Maragos, and F. Lefebvre-Albaret, “The dicta-sign wiki: Enabling web communication for the deaf,” in *Computers Helping People with Special Needs: 13th International Conference, ICCHP 2012, Linz, Austria, July 11-13, 2012, Proceedings, Part II 13*, pp. 205–212, Springer, 2012.
- [86] X. Chai, H. Wang, and X. Chen, “The devisign large vocabulary of chinese sign language database and baseline evaluations,” in *Technical report VIPL-TR-14-SLR-001. Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS)*, Institute of Computing Technology, 2014.
- [87] V. Viitaniemi, T. Jantunen, L. Savolainen, M. Karppa, and J. Laaksonen, “S-pot—a benchmark in spotting signs within continuous signing,” in *LREC proceedings*, European Language Resources Association (LREC), 2014.
- [88] A. Duarte, S. Palaskar, L. Ventura, D. Ghadiyaram, K. DeHaan, F. Metze, J. Torres, and X. Giro-i Nieto, “How2sign: a large-scale multimodal dataset for continuous american sign language,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2735–2744, 2021.
- [89] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze, “How2: a large-scale dataset for multimodal language understanding,” *arXiv preprint arXiv:1811.00347*, 2018.
- [90] I. Grishchenko and V. Bazarevsky, “Mediapipe holistic.” <https://google.github.io/mediapipe/solutions/holistic.html>. Accessed: 15 April 2024.
- [91] A. Moryossef, M. Müller, and R. Fahrni, “pose-format: Library for viewing, augmenting, and handling. pose files,” *arXiv preprint arXiv:2310.09066*, 2023.
- [92] “Sign mediapipe vq.” <https://github.com/sign-language-processing/sign-vq>. Accessed: 15 April 2024.
- [93] J. Zelinka and J. Kanis, “Neural sign language synthesis: Words are our glosses,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3395–3403, 2020.
- [94] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser,

- and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [95] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- [96] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” *arXiv preprint arXiv:1607.06450*, 2016.
- [97] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Advances in neural information processing systems*, vol. 27, 2014.
- [98] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” *arXiv preprint arXiv:1411.1784*, 2014.
- [99] A. L. Maas, A. Y. Hannun, A. Y. Ng, *et al.*, “Rectifier nonlinearities improve neural network acoustic models,” in *Proc. icml*, vol. 30, p. 3, Atlanta, GA, 2013.
- [100] F. Yunus, C. Clavel, and C. Pelachaud, “Sequence-to-sequence predictive model: From prosody to communicative gestures,” in *International Conference on Human-Computer Interaction*, pp. 355–374, Springer, 2021.
- [101] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, “Generating sentences from a continuous space,” *arXiv preprint arXiv:1511.06349*, 2015.
- [102] X. Cheng, W. Xu, T. Wang, and W. Chu, “Variational semi-supervised aspect-term sentiment analysis via transformer,” *arXiv preprint arXiv:1810.10437*, 2018.
- [103] L. Fang, T. Zeng, C. Liu, L. Bo, W. Dong, and C. Chen, “Transformer-based conditional variational autoencoder for controllable story generation,” *arXiv preprint arXiv:2101.00828*, 2021.
- [104] J. Jiang, G. G. Xia, D. B. Carlton, C. N. Anderson, and R. H. Miyakawa, “Transformer vae: A hierarchical model for structure-aware and interpretable music representation learning,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 516–520, IEEE, 2020.
- [105] Z. Lin, G. I. Winata, P. Xu, Z. Liu, and P. Fung, “Variational transformers for diverse response generation,” *arXiv preprint arXiv:2003.12738*, 2020.
- [106] P. Baldi, “Autoencoders, unsupervised learning, and deep architectures,” in *Proceedings of ICML workshop on unsupervised and transfer learning*, pp. 37–49, JMLR Workshop and Conference Proceedings, 2012.
- [107] L. Tarrés, G. I. Gállego, A. Duarte, J. Torres, and X. Giró-i Nieto, “Sign language translation from instructional videos,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5624–5634, 2023.
- [108] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen,

- “Improved techniques for training gans,” *Advances in neural information processing systems*, vol. 29, 2016.
- [109] M. Kipp, A. Heloir, and Q. Nguyen, “Sign language avatars: Animation and comprehensibility,” in *Intelligent Virtual Agents: 10th International Conference, IVA 2011, Reykjavik, Iceland, September 15-17, 2011. Proceedings 11*, pp. 113–126, Springer, 2011.
- [110] J. McDonald, R. Wolfe, J. Schnepf, J. Hochgesang, D. G. Jamrozik, M. Stumbo, L. Berke, M. Bialek, and F. Thomas, “An automated technique for real-time production of lifelike animations of american sign language,” *Universal Access in the Information Society*, vol. 15, pp. 551–566, 2016.
- [111] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.
- [112] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.