Imane Boumedra

AI-Powered Assistive Translator Designed For Deaf & Hard-of-Hearing Individuals

Imane Boumedra

Final International University
June 2026
Girne, North Cyprus

AI-Powered Assistive Translator Designed For Deaf-Mute Individuals

MS, 2026

AI-Powered Assistive Translator Designed For Deaf & Hard-of-Hearing Individuals

by

Imane Boumedra

A thesis submitted to the Institute of Graduate Studies
in partial fulfillment of the requirements for the Degree of
Master of Science
in
Software Engineering

Final International University
June 2026
Girne, North Cyprus

FINAL INTERNATIONAL UNIVERSITY
INSTITUTE OF GRADUATE STUDIES

APPROVAL

Title: AI-Powered Assistive Translator Designed For Deaf & Hard-of-Hearing Individuals

Approval of the Examining Committee:

Asst. Prof. Dr. Payam Zarbakhsh (Chair)

_____ _____

Asst. Prof. Dr. Hüseyin Lort

_____ _____

Asst. Prof. Dr. Ghazaal Sheikhi (Supervisor)

_____ _____

Approval of the Institute of Graduate Studies:

_____

Asst. Prof. Dr. Seniha Besim
Director

Imane Boumedra

imane.boumedra@final.edu.tr

ORCID ID: 0009-0006-4654-1613

©Imane Boumedra 2026

Loving tribute to my dear grandmother. May Allah elevate her rank in Jenna.

Kindness, wisdom, and strong faith were qualities that made her admirable and have greatly influenced my own character. With the decline of her hearing capabilities, I got the opportunity to observe the difficulties that could be caused by hearing problems. Her experiences became the motivation for me to think about ways in which one could make life easier for other people by means of advanced technology.

This is why she was the driving force that inspired me to choose this topic for my thesis. Though she is no longer here, her memory inspires me now. This project is a tribute to her and all the values that she imparted into her children and grandchildren.

ETHICAL DECLARATION

I, Imane BOUMEDRA, hereby, declare that I am the sole author of this thesis and it is my original work. I declare that I have followed ethical standards in collecting and analyzing the data and accurately reported the findings in this thesis. I have also properly credited and cited all the sources included in this work.

Imane Boumedra

ACKNOWLEDGMENTS

I would like to extend my heartfelt gratitude to **Asst. Prof.c Dr. Ghazaal Sheikhi**, my supervisor, who guided me through every step of this project. Her insightful guidance has been critical to the development of this study, from its humble beginning as a Sign Language Transformer base model to a full-fledged experimental study on multilingual sign language translation.

I would like to sincerely thank the Final International University and the faculty of engineering for their support and resources, which made it possible for me to complete my studies at the level of master's degree in software engineering. Additionally, I would like to express my gratitude to the authors of the datasets PHOENIX-2014-T and the implementation of the Sign Language Transformer baseline. Their dedication to open research and their quality data and methods have provided the cornerstone for this work.

On a personal level, this accomplishment is dedicated to my mother, father, my brothers and my sister who provided me with unconditional love and support all along throughout my academic journey. This could not be possible without the unwavering belief they had in my capabilities especially during extended periods of time spent experimenting, training, writing and editing.

To conclude, I would like to thank my friends and peers for the encouragement and company. They were always there for me, ready to listen to me and help out when the going was tough.

Any remaining errors or shortcomings in this work are solely my own responsibility.

ABSTRACT

Accessibility in cross-lingual deployment scenarios is limited because sign language translation (SLT) systems are typically trained for a single spoken-language target. This paper presents a novel two-stage framework that extends monolingual SLT optimisation to a cascaded multilingual text generation pipeline under realistic computational constraints. Building on a Sign Language Transformer baseline, we conduct systematic configuration-driven experimentation from baseline to optimised configuration for English and German over the PHOENIX-2014-T benchmark, identifying a high-performing setting (Optimised configuration–German) that achieves WER 31.98 and BLEU-4 21.23 on the German test set, and WER 48.39 and BLEU-4 20.88 on the English pivot test set. We reconstruct a fully aligned English-supervised dataset from PHOENIX-2014-T release-v3 annotations (train 7,096 / dev 519 / test 642), preserving pre-extracted visual features. Engineering contributions include checkpoint-safe resumable training and robust annotation-alignment logic for heterogeneous formats. The validated English outputs serve as pivot text for a cascaded multilingual extension in which a post-SLT text translation model maps SLT outputs to additional target languages without retraining the visual backbone. A reproducible end-to-end workflow is provided.

Keywords: Sign language translation, PHOENIX- 2014-T, multi-task learning, cascaded translation, pivot language.

ÖZ

İşaret dili çeviri (SLT) sistemleri geleneksel olarak tek bir konuşulan dil hedefi için eğitilir ve bu da diller arası uygulama senaryolarında erişilebilirliği sınırlar. Bu makale, tek dilli SLT optimizasyonunu gerçekçi hesaplama kısıtlamaları altında kademeli çok dilli metin oluşturma hattına genişleten yeni bir iki aşamalı çerçeve sunmaktadır. İşaret Dili Dönüştürücü temel alınarak, PHOENIX-2014-T kıyaslama testi üzerinde İngilizce ve Almanca için temelden optimize edilmiş yapılandırmaya kadar sistematik yapılandırma odaklı deneyler yürütülmekte ve Almanca test setinde WER 31.98 ve BLEU-4 21.23, İngilizce pivot test setinde ise WER 48.39 ve BLEU-4 20.88 değerlerine ulaşan yüksek performanslı bir ayar (Optimize edilmiş yapılandırma - Almanca) belirlenmektedir. Önceden çıkarılmış görsel özellikleri koruyarak, PHOENIX-2014-T sürüm-v3 açıklamalarından (eğitim 7096, geliştirme 519, test 642) tamamen hizalanmış İngilizce denetimli bir veri seti yeniden oluşturuyoruz. Mühendislik katkıları arasında, kontrol noktası güvenli, devam ettirilebilir eğitim ve heterojen formatlar için sağlam açıklama hizalama mantığı yer almaktadır. Doğrulanan İngilizce çıktılar, görsel omurgayı yeniden eğitmeden SLT çıktılarını ek hedef dillere eşleyen bir SLT sonrası metin çeviri modelinin yer aldığı kademeli çok dilli bir uzantı için temel metin görevi görür. Tekrarlanabilir uçtan uca bir iş akışı sağlanmıştır..

Anahtar Kelimeler: İşaret dili çevirisi, PHOENIX-2014-T, çoklu görev öğrenimi, kademeli çeviri, pivot dil.

TABLE OF CONTENTS

ETHICAL DECLARATION vi
ABSTRACT viii
ÖZ ix
LIST OF TABLES xii
LIST OF FIGURES xiii
LIST OF APPENDICES xiv
CHAPTER 1 1
INTRODUCTION 1
CHAPTER 2 3
LITERATURE REVIEW 3
2.1 Accessibility Employing Deployed Systems and Gesture-Level Recognition 3
2.2 Sign Language Recognition and Translation 4
2.3 Multilingual Neural Machine Translation 4
2.4 Cascaded vs. End-to-End Architectures 5
2.5 Pre-trained Language Models For Gloss-to-text Translation 5
2.6 Open Vocabulary Sign Language Translation (OVSLT) 6
2.7 Recent SLT and Multilingual Extensions 6
CHAPTER 3 8
METHODOLOGY 8
3.1 Model Architecture and Data 8
3.1.1 Sign Language Transformer Backbone 8
3.1.2 Visual Features 11
3.1.3 Dataset: PHOENIX-2014-T 11
3.1.4 Justification of Baseline Selection 11
3.2 Optimization Driven by Configuration (from baseline to ablation) 12
3.2.1 Motivation and Design 12
3.2.2 Configuration Variants and Intermediate Results 13
3.2.3 Selected Configuration 13
3.2.4 Training Engineering: Resumable Checkpointing 14
3.3 English Dataset Reconstruction 14
3.3.1 Motivation 14
3.3.2 Alignment Results 15
3.4 Training Objective and Evaluation Metrics 15
3.4.1 Joint Multi-Task Objective 15
3.4.2 Recognition Metric: Word Error Rate 16
3.4.3 Translation Metric: BLEU 16
3.4.4 Decoding Configuration 16

CHAPTER 4 17
RESULTS 17
4.1 Experimental Results 17
4.1.1 Training Overview 17
4.1.2 Prior Work Baseline (German Test Set) 30
4.1.3 Unified Multilingual DE+EN Results (Shared Model) 30
4.1.4 Discussion Of The Current Bottleneck 33
4.2 Cross-Lingual Cascade Extension 35
4.2.1 Protocol 35
4.2.2 Pilot Translation Results (EN $\rightarrow$ FR) 36
4.2.3 Discussion 36
CHAPTER 5 38
CONCLUSION AND IMPLICATIONS 38
5.1 Conclusion 38
5.2 Future Work 39
5.3 Data and Code Availability 39
REFERENCES 40
APPENDICES 43

LIST OF TABLES

Table 1 Summarizes Intermediate Training Metrics and Key Changes Between Versions. 13
Table 2 English Dataset Reconstruction Alignment Statistics. 15
Table 3 Baseline Results on PHOENIX-2014-T (German Test Set). 30
Table 4 Unified Multilingual DE+EN Model Results. 31
Table 5 Comparison of Monolingual and Multilingual SLT Models. 32
Table 6 Per-language DE+EN-v4 Metrics on The PHOENIX-2014-T Test Set. . . 32
Table 7 Pilot Cascade Results: English Pivots Generated By DE+EN-v4 and Translated to French Using NLLB. 36

LIST OF FIGURES

Figure 1 Unified Multilingual SLTR (DE+EN) With Language-specific Translation Heads. Shared Encoder and Decoder Trunk, One Shared CTC Gloss Head, and Separate Final Linear Layers For <de> and <en> Translation. 9
Figure 2 Joint Sign Language Transformer Baseline (Sign2Gloss + Text). The bilingual Variant Adds Language-specific Translation Heads For Joint DE+EN Training. 10
Figure 3 V1 Loss Curves 18
Figure 4 V1 Recognition Metrics 19
Figure 5 V1 Translation Metrics 20
Figure 6 V2 Loss Curves 21
Figure 7 V2 Recongnition Metrics 21
Figure 8 V2 Translation Metrics 22
Figure 9 V3 Loss Curves 23
Figure 10 V3 Recongnition Metrics 23
Figure 11 V3 Translation Metrics 24
Figure 12 V4 Loss Curves 25
Figure 13 V4 Recongnition Metrics 26
Figure 14 V4 Translation Metrics 27
Figure 15 V4-Engl Loss Curves 28
Figure 16 V4-Engl Recongnition Metrics 29
Figure 17 V4-Engl Translation Metrics 29
Figure 18 Error Propagation in The Two-stage Cascaded Sign-to-target Translation Pipeline. 34
Figure 19 End-to-end Multilingual Cascade Framework. An Unified DE+EN SLTR model. 35

LIST OF APPENDICES

APPENDIX A: Similarity Report 43
APPENDIX B: Statement of Artificial Intelligence Usage 50

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
API	Application Programming Interface
ASL	American Sign Language
BLEU	Bilingual Evaluation Understudy
BP	Brevity Penalty (BLEU component)
BPN	Backpropagation Neural Network
BSL	British Sign Language
chrF	Character *n*-gram F-score
CLIP	Contrastive Language–Image Pre-training
CNN	Convolutional Neural Network
COMET	Crosslingual Optimized Metric for Evaluation
CorrNet	Correlation Network (continuous sign language recognition)
CSL	Chinese Sign Language
CSL-Daily	Chinese Sign Language Daily dataset
CSLR	Continuous Sign Language Recognition
CTC	Connectionist Temporal Classification
CVPR	IEEE/CVF Conference on Computer Vision and Pattern Recognition
DE	German (Deutsch)ö language tag `<de>`
DGS	German Sign Language (Deutsche Gebärdensprache)
DHH	Deaf and Hard-of-Hearing
EN	Englishö language tag `<en>`
FIU	Final International University
FLORES	Facebook Low Resource (language-code scheme used by NLLB)
FR	French
GFSLT-VLP	Gloss-Free Sign Language Translation with Visual-Language Pretraining
GPU	Graphics Processing Unit
HOG	Histogram of Oriented Gradients
ICCV	IEEE/CVF International Conference on Computer Vision
JIIS	Journal of Intelligent Information Systems
LR	Learning Rate
METEOR	Metric for Evaluation of Translation with Explicit ORdering
ML	Machine Learning
MLE	Maximum Likelihood Estimation
MT	Machine Translation
mBART	Multilingual BART (denoising sequence-to-sequence pre-training)
mT5	Massively Multilingual T5 (text-to-text transformer)

NAACL	North American Chapter of the Association for Computational Linguistics
NLLB	No Language Left Behind (multilingual neural machine translation)
NMT	Neural Machine Translation
ODE	Optimised German–English configuration (Model-ODE, German v4)
OEN	Optimised English configuration (Model-OEN)
ORCID	Open Researcher and Contributor ID
OVSLT	Open-Vocabulary Sign Language Translation
PHOENIX	Public benchmark corpus family (PHOENIX-2014, PHOENIX-2014-T)
PLM	Pre-trained Language Model
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
RNN	Recurrent Neural Network
RWTH	Rheinisch-Westfälische Technische Hochschule Aachen
sEMG	Surface Electromyography
SLP	Sign Language Production
SLR	Sign Language Recognition
SLT	Sign Language Translation
SLTR	Sign Language Translation and Recognition (joint gloss + text model)
SMKD	Self-Mutual Knowledge Distillation
SOTA	State of the Art
T5	Text-to-Text Transfer Transformer
TACL	Transactions of the Association for Computational Linguistics
TX	Translation (beam-search branch in decode configuration)
VLP	Visual-Language Pretraining
WCAG	Web Content Accessibility Guidelines
WER	Word Error Rate
WHO	World Health Organization

CHAPTER 1
INTRODUCTION

Sign language translation (SLT) is a sequence-to-sequence task that converts continuous sign language video into spoken-language text. Existing systems typically target a single spoken language, thereby restricting deployment in multilingual contexts and limiting the reach of accessibility technology for deaf communities worldwide. Scaling SLT to multiple languages poses a fundamental challenge: full end-to-end retraining for every target language is computationally prohibitive, particularly when visual feature extraction is expensive. A feasible solution would be to implement a cascading approach where a solid bilingual SLT core such as DE and EN provides input to a lightweight multilingual translation model layer that offers wide language coverage without having to retrain the visual component.

This work has five main contributions:

1. Two-phase cascaded SLT pipeline architecture leveraging bilingual SLT backbone (DE/EN) and a separate multilingual text-translation stage to support extra target languages without having to retrain the visual backbone

2. An English-labeled, fully aligned corpus based on PHOENIX-2014-T Release-V3 for standardized cross-lingual SLT evaluation

3. Empirical analysis of recognition/translation trade-offs in joint multi-task SLT transformers with shared training and decoding

4. Evaluation protocol including error decomposition reporting and decode-grid model selection (CTC beam + translation beam/alpha)

5. Translation output layers depending on “de”/“en” tokens (DE+EN-v4) maintaining one CTC gloss layer while substantially increasing the translation performance among languages.

The rest of the thesis is organized as follows. Chapter 2 covers the background of related work done in sign language recognition, sign language translation, multilingual translation, and cascaded systems. Chapter 3 explains the proposed approach by discussing the method, the architecture of the proposed model, the preprocessing steps applied, configuration analysis, English data restoration, and measures utilized for evaluation. In chapter 4, experimental results are discussed covering training dynamics, analysis of the proposed unified system including the DE+EN models along with their respective experimental results, and finally cross-language cascade system analysis. Finally, chapter 5 discusses the conclusions derived from the entire thesis and future work directions.

CHAPTER 2
LITERATURE REVIEW

2.1 Accessibility Employing Deployed Systems and Gesture-Level Recognition

The need for communication is innate to human beings; however, a large number of the population struggles because of hearing disabilities and speech disorders. As per the statistics of World Health Organization World Health Organisation (2024), roughly 430 million individuals around the world, comprising 34 million children, need rehabilitation services owing to their severe hearing loss, constituting more than 5% of the global population.. This number is predicted to increase to nearly 700 million by 2050, which means that one in ten people will be impacted.

The Deaf and hard-of-hearing communities employ sign languages, which are visual-gestural communication systems. Sign language recognition (SLR) is at the nexus of technology and accessibility, as noted Harshitha, Rajagopalan, et al. (2024). Automatic recognition and translation can lower communication barriers in public services, education, employment, and healthcare.

It is still challenging to facilitate efficient communication between signers and non-signers. Interaction frequently depends on lipreading and sign language, which are skills that need advanced training and are not equally available to everyone because of time, money, and the scarcity of specialized teaching.

Over the past decade, sign language recognition has advanced dramatically from traditional image processing to machine learning pipelines (preprocessing, feature extraction, classification). The efficacy of Backpropagation Neural Networks (BPN) for sign language identification utilizing Histogram of Oriented Gradients (HOG) features was shown by Sreemathy et al. (2023). For ASL recognition,Singh

and Chaturvedi (2023) investigated surface electromyography (sEMG) sensors using machine learning methods. In isolated or controlled environments, deep learning techniques, such as CNNs for spatial features, have demonstrated impressive results (e.g., Obi et al. (2023), Mohsin et al. (2023) on transfer learning, and Nandi et al. (2023) on architectures for Indian Sign Language).

Commercial and assistive technologies include wearables like sensorized gloves Babour et al. (2023), neural sign-language translation pipelines Attar et al. (2022), sign-language products like SignAll for ASL Papatsimouli et al. (2023), and speech-to-text systems (like AVA, Google Live Transcribe, and Microsoft Translator) that support spoken language but not sign video. High stated accuracy frequently results from controlled conditions rather than in-the-wild signing, and many systems prioritize one-way sign-to-text or sign-to-speech help over symmetric interaction. The complementing regime of joint recognition and translation from continuous video characteristics is addressed by continuous sentence-level SLT on benchmarks like PHOENIX-2014-T (section 3.1.3).

2.2 Sign Language Recognition and Translation

Sign language recognition (SLR) is the task of converting a sequence of frames of sign video to gloss sequences, while SLT maps video directly to spoken-language sentences. Camgoz et al. (2018) introduced the PHOENIX-2014-T benchmark and a CNN–RNN baseline for continuous SLT. Their subsequent Transformer-based model (Camgoz et al., 2020) established the state of the art and provides the backbone for the present work, framing recognition and translation as a joint multi-task objective. More recent approaches incorporate attention-based spatial modelling (Li et al., 2020) and large pre-trained visual representations (**zuo2022**).

2.3 Multilingual Neural Machine Translation

Multilingual NMT has demonstrated that a single model can translate across many language pairs when conditioned on a language token (Johnson et al., 2017).

Massively multilingual pre-trained models such as mBART (Liu et al., 2020) and mT5 (Xue et al., 2021) provide strong initialisation for low-resource translation settings. Our cross-lingual extension leverages such models as a second-stage module operating on SLT pivot text, avoiding the need to retrain visual components per language.

2.4 Cascaded vs. End-to-End Architectures

Cascaded architectures decompose complex tasks into modular stages, trading flexibility for robustness and scalability (Sperber & Paulik, 2020). In SLT, a cascaded sign→gloss→text approach has been shown to outperform direct sign-to-text systems when gloss annotations are available (Camgoz et al., 2020). Our two-stage cascaded multilingual design extends this principle by inserting a language-routing translation module after the SLT output, enabling broader language coverage at reduced computational cost.

2.5 Pre-trained Language Models For Gloss-to-text Translation

By fine-tuning big pre-trained language models (PLMs), such as T5, Flan-T5, mBART, and Llama, on several SLT-related corpora, such as RWTH-PHOENIX-2014T, (Kermani et al., 2025) study **bidirectional gloss–text translation**. Their study demonstrates notable improvements over Transformer baselines built from the ground up on a range of widely used machine translation parameters.

However, the present work differs greatly from their notion. In particular, Kermani et al. work in a **text-to-text** environment where the translation model receives *gloss sequences as input*. On the other hand, our method tackles the joint **video-feature → gloss + text** SLT problem, in which a unified multi-task Transformer architecture is used to simultaneously learn recognition and translation from continuous sign-language visual data.

Accordingly, this work is not used as a direct numerical comparator to **Model-ODE** or the primary SLT results reported in our experiments. Instead, it provides supporting motivation for the proposed **PLM-based second-stage cascade**

extension discussed in Section 8, where multilingual text generation models may be integrated downstream of the SLT pivot output to improve cross-lingual scalability.

2.6 Open Vocabulary Sign Language Translation (OVSLT)

(Wang et al., 2025) introduce **OVSLT**, an open-vocabulary sign language translation framework designed to address the vocabulary limitations of conventional gloss-based SLT systems. Their approach emphasises semantic descriptions and grid-based visual–textual alignment, and is evaluated on benchmark datasets including PHOENIX-2014-T and CSLDaily.

The OVSLT framework differs substantially from the present work in terms of architectural design, supervision strategy, and evaluation emphasis. In particular, their formulation focuses on open-vocabulary semantic alignment and reports performance using a broader set of generation-oriented metrics, including ROUGE, whereas our study centres on joint gloss recognition and translation optimisation using WER and BLEU-4 as primary evaluation criteria.

Accordingly, we position OVSLT as **recent complementary work** addressing representation and vocabulary constraints in classical gloss-based SLT pipelines, rather than as a directly comparable baseline to our multilingual Transformer configurations. For this reason, we avoid merging OVSLT headline results into the main WER/BLEU comparison tables unless accompanied by a dedicated "related systems" comparison row explicitly specifying task formulation, supervision type, input modality, and metric definitions to ensure fair interpretation across heterogeneous SLT settings.

2.7 Recent SLT and Multilingual Extensions

Several recent advancements have been made surpassing the baseline performance set by Camgoz et al. (2020) on the PHOENIX-2014-T dataset, especially regarding gloss recognition accuracy and translation quality. These advancements usually involve either changes in the visual feature extraction pipeline, multimodal representation learning, or training objective. For instance, CorrNet Hu et al.

(2023) proposes correlation-based learning strategies facilitating greater interaction between visual and linguistic representations. Sign Language Multi-task Knowledge Distillation (SMKD) by Chen et al. (2022) utilizes sign language multi-task knowledge distillation for enhanced recognition and translation quality, whereas GFSLT-VLP Zhou et al. (2023) exploits large-scale vision-language pre-training for better cross-modal representations. Despite offering remarkable benchmarks for performance, such techniques necessitate retraining or replacement of the visual feature extractor and hence are not comparable to the fixed 1024-dimensional PHOENIX features used in this thesis.

Simultaneously, developments in multilingual neural machine translation allow for efficient scaling from one source language to many target languages without changing the input model. The availability of large-scale multilingual sequence-to-sequence models, including but not limited to mBART Liu et al. (2020), mT5 Xue et al. (2021), and NLLB "Scaling neural machine translation to 200 languages" (2024), allows for high-quality pivot translation between many languages. Instead of attempting to achieve the best possible PHOENIX-2014-T speech recognition results, this thesis chooses a different angle for research, focusing on the extension of a solid reproducible baseline for SLT with English supervision through joint German and English training and a language-specific decoding head, as well as implementing a multilingual cascade system. These decisions make this work more feasible and transparent while preparing for further development involving more powerful visual models and multilingual translation systems

CHAPTER 3
METHODOLOGY

3.1 Model Architecture and Data

3.1.1 Sign Language Transformer Backbone

Our system is built on the `signjoey` implementation of the Sign Language Transformer (Camgoz et al., 2020). The architecture is a multi-task encoder–decoder Transformer with two output branches:

1. CTC-based recognition branch producing gloss sequences
2. A cross-attention translation branch producing spoken-language text.

For unified DE+EN training, targets are prefixed with language tags *<de>* and *<en>* so that the decoder attends to conditioning tokens in the target sequence.

Multilingual refinement (DE+EN-v4):

We add **language-specific translation linear layers** to the decoder in addition to a single shared vocabulary softmax: one head is activated when the target language tag is <de>, and another when it is <en>. Only the final logits projection before loss computation and beam search is separated, while the Decoder Transformer layers and embeddings remain shared. Although spoken-language outputs improve, tag or mixed-language gloss errors may still occur because gloss recognition is still modeled using a **single shared** CTC head over the combined gloss vocabulary.

Figure 1

Unified Multilingual SLTR (DE+EN) With Language-specific Translation Heads. Shared Encoder and Decoder Trunk, One Shared CTC Gloss Head, and Separate Final Linear Layers For <de> and <en> Translation.

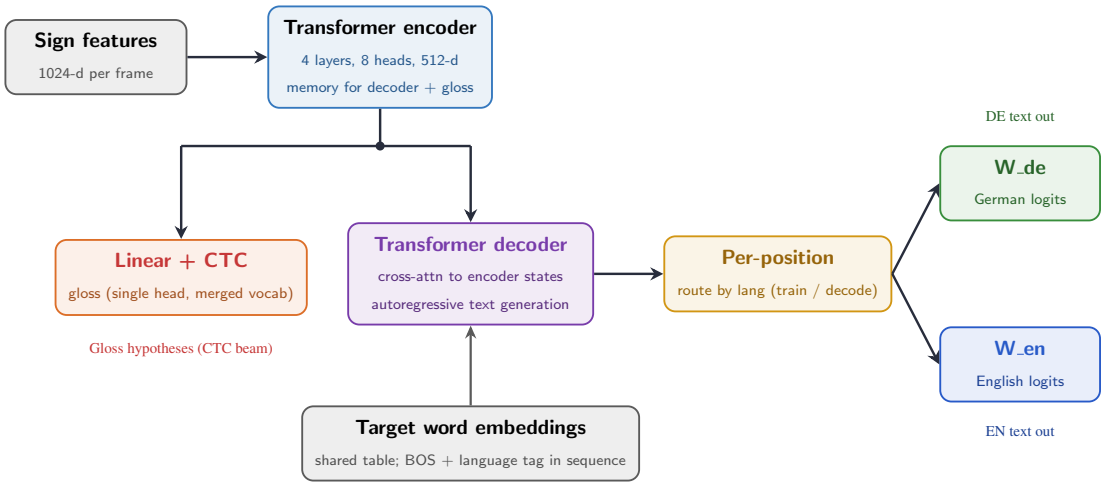


Figure 1 shows the DE+EN-v4 model. Sign features (1024-dimensional vectors per frame) are processed by a shared Transformer encoder. Two branches share this encoder.

The **recognition branch** uses Connectionist Temporal Classification (CTC) to predict gloss sequences (sign-language words). A single shared gloss prediction head is used for both German and English. The gloss head is not split by language, so recognition errors may still propagate across languages.

The **translation branch** uses a shared Transformer decoder to generate spoken-language text. Target sequences begin with language tags (<de> or <en>) indicating the desired output language. In DE+EN-v4, the final output projection layer before the vocabulary is split into separate German and English heads, while all lower decoder layers remain shared. This design reduces interference between German and English text generation while preserving shared multilingual representations.

The model is trained with a joint objective consisting of recognition loss and translation loss. By combining a shared visual encoder with language-specific output heads, DE+EN-v4 maintains a single visual model while improving language-specific text generation.

The joint loss is:

$$\mathscr{L}_{\text{total}} = \lambda_{\text{rec}} \cdot \mathscr{L}_{\text{rec}} + \lambda_{\text{tr}} \cdot \mathscr{L}_{\text{tr}}, \tag{1}$$

where $\mathscr{L}_{\text{rec}}$ is the CTC recognition loss over gloss targets and $\mathscr{L}_{\text{tr}}$ is the cross-entropy translation loss over text targets.

For the optimised configuration, $\lambda_{\text{rec}} = 2.5$ and $\lambda_{\text{tr}} = 1.0$. Both encoder and decoder use 4 Transformer layers, 8 attention heads, hidden dimension 512, feed-forward dimension 2048, and dropout 0.1.

DE+EN-v4 fine-tuning. The language-head model is fine-tuned from the best DE+EN-v3 checkpoint using non-strict loading so that newly added head parameters can be initialized and trained. The experiment configuration uses $\lambda_{rec} = 2.2$, $\lambda_{tr} = 1.2$, learning rate 3×10^{-4}, batch size 48, plateau scheduling (patience 12, decay 0.8), label smoothing 0.1, and early stopping based on validation BLEU.

Figure 2

Joint Sign Language Transformer Baseline (Sign2Gloss + Text). The bilingual Variant Adds Language-specific Translation Heads For Joint DE+EN Training.

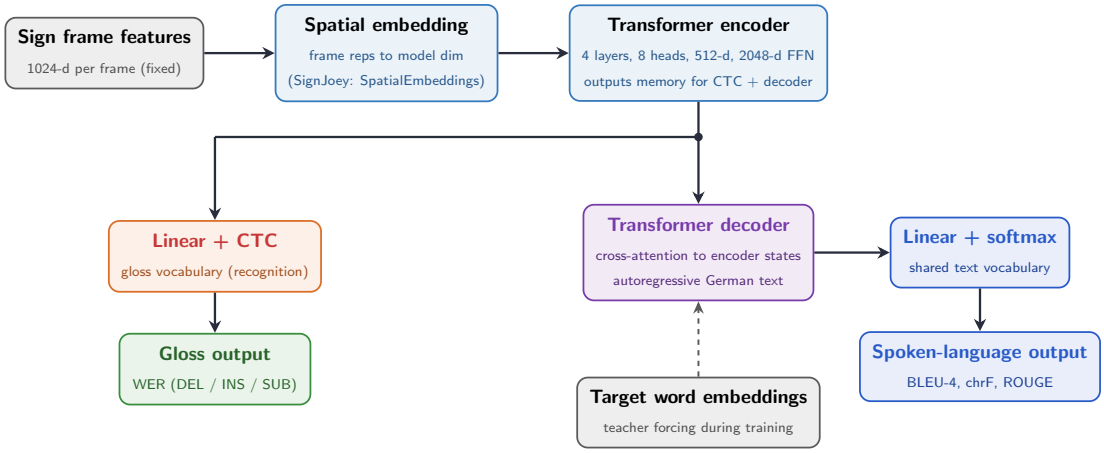


Figure 2 illustrates the standard joint Sign Language Transformer architecture introduced by Camgoz et al. (2020). Processing of visual sign feature representations takes place using one common Transformer encoder. The encoded representation is utilized by two different tasks, with the recognition task applying CTC in order to perform predictions of gloss sequences and thus obtain the target function for sign language recognition.

In parallel, a translation task operates via a Transformer decoder, which directly translates encoded sign features into the respective speech-language sentences.

Both tasks are simultaneously learned with a common visual encoder, enabling representation learning with respect to sign recognition and sign language translation at the same time. This model structure forms a basis for Model-ODE and Model-OEN models.

The DE+EN-v4 architecture shown in Figure 1 extends this baseline by introducing language tags and language-specific translation output heads for bilingual German–English training while retaining a shared visual encoder and shared lower decoder layers.

3.1.2 Visual Features

Frame-level sign representations are drawn from pre-extracted feature files with feature dimension 1024. This work does *not* re-extract visual features from raw video; the visual backbone is treated as fixed across all experiments. English-supervised experiments, therefore, share identical visual inputs with German-supervised runs, modifying only the gloss and text supervision labels.

3.1.3 Dataset: PHOENIX-2014-T

All experiments use the PHOENIX-2014-T dataset, a continuous German Sign Language corpus derived from weather-broadcast footage. Standard data splits are: train (7,096 samples), dev (519), and test (642). English annotations are sourced from the PHOENIX-2014-T release-v3 annotation files (Section 3.3).

3.1.4 Justification of Baseline Selection

The Sign Language Transformer Camgoz et al. (2020) is adopted as the joint SLT backbone in this thesis for four primary reasons:

1. **Reproducibility:** The PHOENIX-2014-T dataset provides publicly available 1024-dimensional frame-level features together with standardized

training, development, and test splits. Using the published Sign Language Transformer baseline enables direct comparison with reported benchmark results, including a German WER of 26.16% and a BLEU-4 score of 21.32 under a well-documented architecture.

2. **Controlled Extension:** By keeping the visual input representation fixed, the study isolates the effects of **language supervision** (German, English, and unified German–English training) and **decoder design** (shared versus language-specific translation heads). This avoids confounding factors that would arise from simultaneously modifying the visual encoder.

3. **Computational Scope:** Recent SLT architectures such as CorrNet and GFSLT-VLP achieve stronger performance through enhanced visual modeling and large-scale pre-training. However, reproducing and retraining such systems would require substantially greater computational resources and development effort than were available within the time and GPU budget of this one-year Master's research project.

4. **Software Engineering Traceability.** The Sign Language Transformer framework supports reproducible experimentation through version-controlled YAML configuration files, checkpoint management, and standardized `signjoey` training and evaluation scripts. These characteristics align well with the requirements for transparency, maintainability, and auditability.

3.2 Optimization Driven by Configuration (from baseline to ablation)

3.2.1 Motivation and Design

Prior to implementing multilingual supervision, we used five versioned configurations (baseline to ablation) assessed on the German PHOENIX-2014-T benchmark to characterize the sensitivity of the SLT backbone to important hyperparameters. Batch size, model depth, learning rate and scheduler strategy, recognition–translation loss balance, and label smoothing were the tuning dimensions

investigated. Finding a reliable, high-performing base configuration prior to English and multilingual transfer was the aim.

3.2.2 Configuration Variants and Intermediate Results

Table 1
Summarizes Intermediate Training Metrics and Key Changes Between Versions.

Version	Key changes	Best WER ↓	Best BLEU-4 ↑	Status
Model-B	$\lambda_{rec} = 2.0$, 3-layer encoder/decoder, learning rate 10^{-3}, batch size 32	84.36	13.16	Exploratory
Model-D	4-layer encoder/decoder (capacity increase); other settings same as baseline	— (deletion-heavy)	$\sim$ 13.09	Exploratory
Model-ODE	$\lambda_{rec} = 2.5$, learning rate 8×10^{-4}, patience 10, decay 0.8, label smoothing 0.1, 4-layer	Dev: 32.03 / Test: 31.98	Dev: 20.62 / Test: 21.23	Best DE model
Model-OEN	Same optimised configuration; English supervision (release-v3); decoding: CTC beam 8, translation beam 4, $\alpha = 2$	Dev: 47.52 / Test: 48.39	Dev: 21.09 / Test: 20.88	Best EN model
Model-A (ablation)	Recognition weight returned to 2.0; other optimised scheduler settings retained	Dev: 35.36 / Test: 35.76	Dev: 20.41 / Test: 20.34	Complete

3.2.3 Selected Configuration

The strongest practical training setup was determined to be the optimized configuration.The most stable convergence behavior and the best overall balance between recognition and translation performance were obtained by combining a higher

recognition lossweight ($\lambda_{rec} = 2.5$), a lower but more steady learning rate (8×10^{-4}), an adaptive plateau scheduler (patience10, decay factor0.8), and label smoothing of 0.1. Model-A, another ablation experiment, kept the existing optimization settings while lowering the recognition weighting to $\lambda_{rec} = 2.0$. Model-A continuously performed worse on both development and test sets than the optimized German configuration **(Model-ODE)** following decode adjustments. **Model-A** achieved BLEU-4 of **20.34** and WER of **35.76** on the test set, whereas **Model-ODE** achieved BLEU-4 of **21.23** and WER of **31.98**.These findings demonstrate that better gloss alignment directly improves downstream translation quality and overall training stability, supporting the use of stronger recognition supervision in the joint multi-task objective.

3.2.4 Training Engineering: Resumable Checkpointing

Long training runs required strong checkpoint management. There were two engineering adjustments made. In order to eliminate purported failures upon restart, the model directory initialization was first modified to minimize existing directories while loading a checkpoint (`overwrite=false`).In order to regulate whether the previous optimizer and scheduler state is carried forward or deleted on resume, explicit reset flags (`resetbestckpt`), (`resetscheduler`), and (`resetoptimizer`) were implemented. These adjustments are necessary for repeatable long-run studies and are recorded in the distributed configuration files.

3.3 English Dataset Reconstruction

3.3.1 Motivation

We created an English-supervised training set using PHOENIX-2014-T release-v3 to enable multilingual output from a shared visual backbone. The goal was to generate annotation files in precisely the binary format that the signjoey training pipeline uses, substituting English equivalents for German gloss and text supervision labels while maintaining the pre-extracted visual characteristics.

Engineering attention was needed for a number of practical issues. A backup chain, `corpus_eng` → `corpus-eng` → `corpus_english` → `corpus.csv`, was used to automate filename discovery. By automatically identifying comma, pipe, and tab delimiters with secure UTF-8 BOM stripping, delimiter and BOM handling was strengthened. In order to address format discrepancies between feature file IDs and annotation IDs, candidate-based matching was used to implement sample identifier normalization.

3.3.2 Alignment Results

The reconstruction pipeline achieved 100% split alignment with zero unmatched samples, as summarised in Table 2.

Table 2
English Dataset Reconstruction Alignment Statistics.

Split	Matched / Total	Size (MB)	Output file
Train	7,096 / 7,096	1,845.56	`phoenix14t_english.train`
Dev	519 / 519	124.77	`phoenix14t_english.dev`
Test	642 / 642	144.31	`phoenix14t_english.test`

3.4 Training Objective and Evaluation Metrics

3.4.1 Joint Multi-Task Objective

Equation (1) defines the weighted total of recognition and translation losses, which the model minimizes. A CTC loss across gloss sequences is called recognition loss $\mathscr{L}_{\text{rec}}$. Cross-entropy loss over target text tokens is known as translation loss $\mathscr{L}_{\text{tr}}$. Stronger recognition supervision stabilizes early training before translation quality can increase, as demonstrated by the optimized setup with $\lambda_{\text{rec}} = 2.5$ and $\lambda_{\text{tr}} = 1.0$.

3.4.2 Recognition Metric: Word Error Rate

Recognition quality is measured by Word Error Rate (WER):

$$\mathrm{WER} = \frac{S + D + I}{N} \times 100, \tag{2}$$

where N is the reference token count and S, D, and I are substitutions, deletions, and insertions with respect to the reference gloss sequence. For every run that is finished, WER decomposition into DEL, INS, and SUB components is supplied, offering diagnostic insight into the distribution of error types.

3.4.3 Translation Metric: BLEU

Translation quality is measured by BLEU-4 (Papineni et al., 2002), computed from modified n-gram precisions ($n = 1, \ldots, 4$) with brevity penalty (BP):

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\frac{1}{4}\sum_{n=1}^{4} \log p_n\right). \tag{3}$$

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{1-r/c} & \text{if } c \leq r \end{cases} \tag{4}$$

BLEU-1 through BLEU-3 are reported as supporting context.

ChrF is additionally reported for the English run as a complementary character-level metric.

3.4.4 Decoding Configuration

CTC beam search and translation beam search are applied independently with tunable beam sizes and length penalty α. For the optimised English pivot model, the best-performing decoding configuration on the test set was CTC beam size 8, translation beam size 4, and $\alpha = 2$. Beam size 1 was used for intermediate checkpoint monitoring during training.

CHAPTER 4

RESULTS

4.1 Experimental Results

4.1.1 Training Overview

Figures 3–17 illustrate model behaviour throughout training for all experimental configurations (V1, V2, V3, V4, and V4-English). For each model version, three plots are provided:

1. Overall training and validation loss,
2. Recognition performance measured by Word Error Rate (WER)
3. Translation performance measured by BLEU-4.

All curves are computed on the **development set** using a simple decoding configuration with beam size 1 during training. Under a constant evaluation environment, these charts are meant to display optimization dynamics, convergence behavior, and relative differences between model variations. The final results reported in Tables 3–5 are obtained using **test-set evaluation** with decoding hyperparameters selected on the development set. In particular, larger beam sizes are used for the final recognition and translation decoders, yielding stronger performance than the beam-1 development curves shown in Figures 3–17. Consequently, the values in the training plots should be interpreted as indicators of training progress rather than as the final scores reported for each model.

Figure 3

V1 Loss Curves

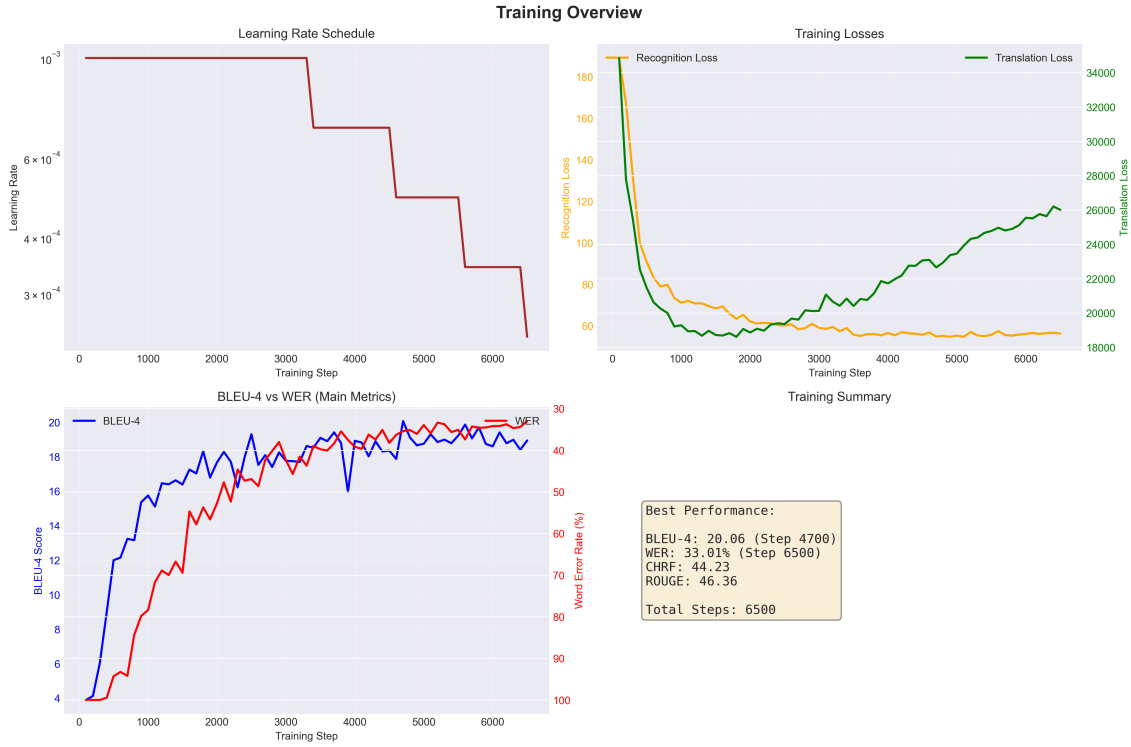


Figure 3 shows the training behaviour of the V1 model. The top-left panel plots the learning rate schedule, which decreases when development-set BLEU no longer improves. The top-right panel shows the two training objectives: the gloss recognition loss (orange) and the translation loss (green). Both losses are expected to decrease as training progresses.

Bottom left panel shows results for the development set, where the score of BLEU-4 (in blue) needs to rise, while Word Error Rate (in red) needs to fall.

In case of V1, the former rises steadily but the latter stays relatively high through all the training, reflecting poor performance on the task of recognizing words. The last two panels represent best scores attained in training on the development set.

Overall, V1 does not learn well and performs poorly at gloss recognition.

Figure 4

V1 Recognition Metrics



Figure 4 presents an analysis of gloss recognition quality in the V1 architecture. Quality is evaluated in terms of Word Error Rate (WER), where smaller values mean better predictions.

Apart from WER itself, this graph also provides its individual elements, including deletions, insertions, and substitutions. Deletion errors occur when the gloss token exists in the reference gloss but does not appear in the prediction output. Insertions are gloss tokens that do not exist in the reference gloss but appear in the prediction output, while substitutions are incorrectly predicted gloss tokens. Deletions have the highest share of errors among all three error components and dominate through the whole learning process, which means that the model does not correctly recognize gloss tokens and output them.

Insertions and substitutions, in contrast, contribute significantly less. Thus, the model is likely to underestimate the number of tokens rather than make any excess or wrong predictions. Deletion errors justified further design changes in terms of deeper networks and higher recognition loss weighting.

Figure 5

V1 Translation Metrics

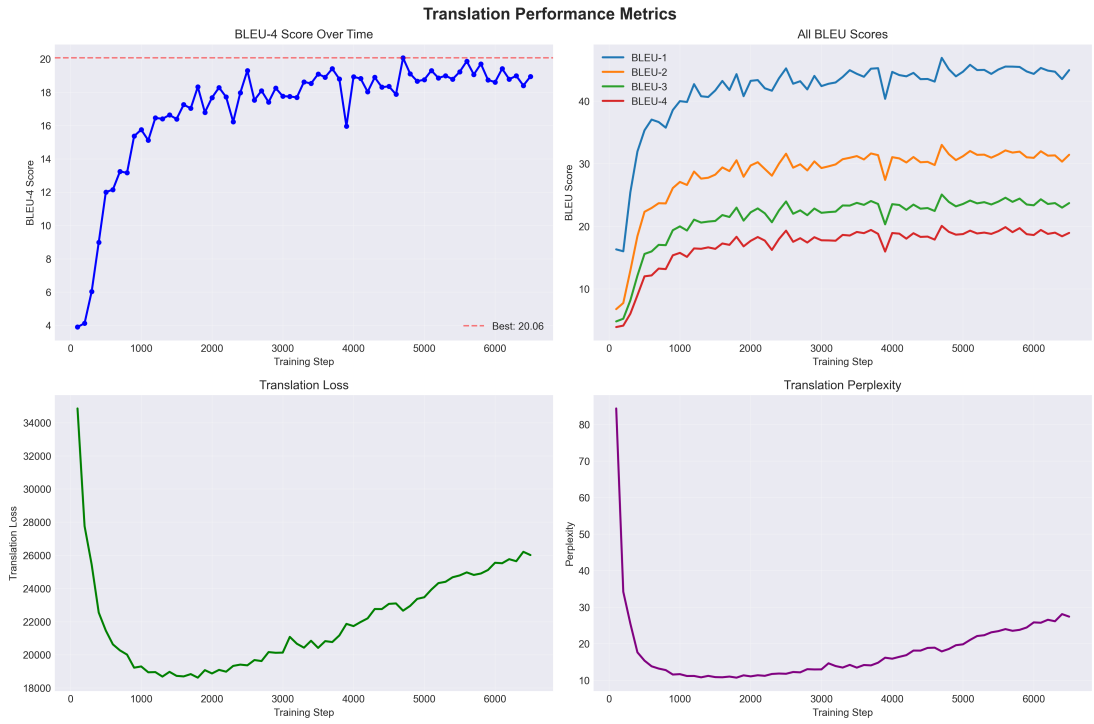


Figure 5 illustrates the quality of translation of the V1 model with respect to spoken language.

The measure used here is BLEU-4, with higher scores reflecting better agreement between the output translations and the sentences given as references. While during training, the BLEU-4 score grows consistently to a rather good translation level, the gloss recognition ability does not develop at all, and stays weak, with high WER scores.

This combination of results implies that the model is able to learn grammatical structures and translation patterns without achieving proper gloss alignment.

This kind of behavior is characteristic of joint Sign Language Transformer models, in which translation and recognition goals are known to be in conflict with each other.

Increased translation performance doesn't mean improved gloss recognition ability.

Figure 6

V2 Loss Curves

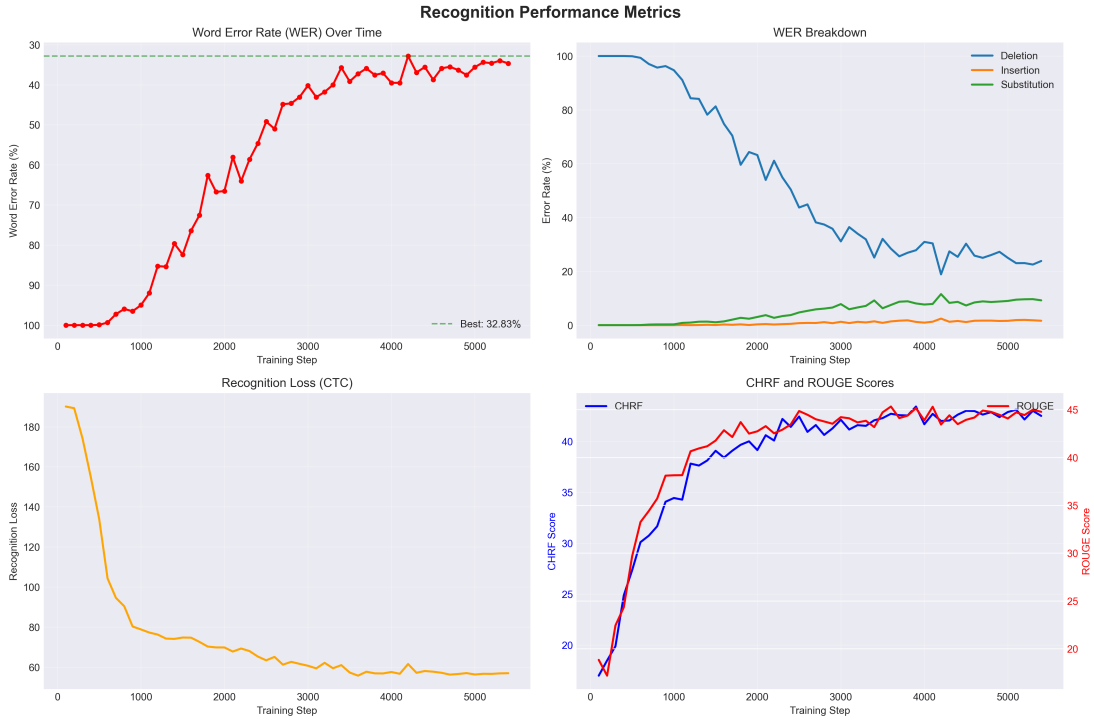


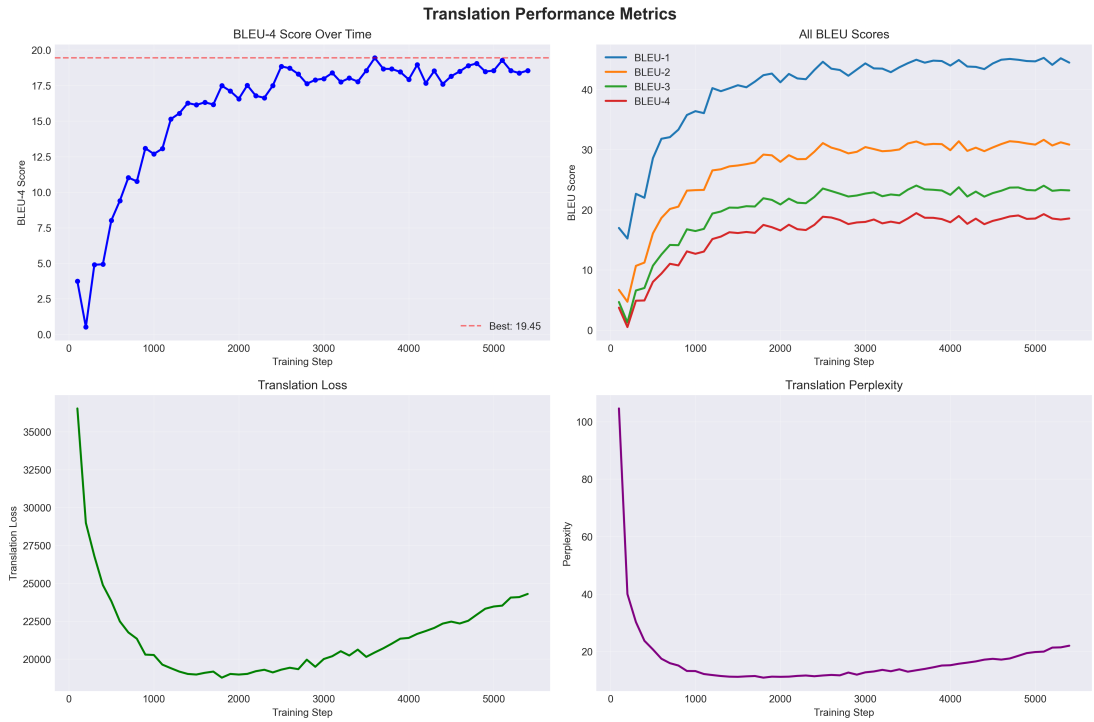
Figure 7

V2 Recongnition Metrics



Figure 8

V2 Translation Metrics



Figures 6–8 illustrate the training behaviour of the V2 model. In comparison to V1, the depth of the Transformer encoder and decoder layers in V2 is increased by one layer each. With added model capacity, V2 tends to outperform V1 in terms of lower WER and higher BLEU-4 scores on the development set.

Based on the results of this improvement in recognition accuracy, we hypothesize that the deeper encoder is better at aligning sign-language features with gloss sequences, leading to fewer recognition errors.

This added capacity helps with translation, as well, judging by the BLEU-4 score increase. It can therefore be concluded that increased depth is beneficial for learning visual-to-gloss and visual-to-text translations. Given that the addition of encoder/decoder layers helped to improve recognition and translation, further exploration with this modification was done in all subsequent models.

However, V2 should not be considered the optimized architecture; it was an intermediate step that allowed us to develop further into the V4 model, which became the basis for Model-ODE

Figure 9

V3 Loss Curves



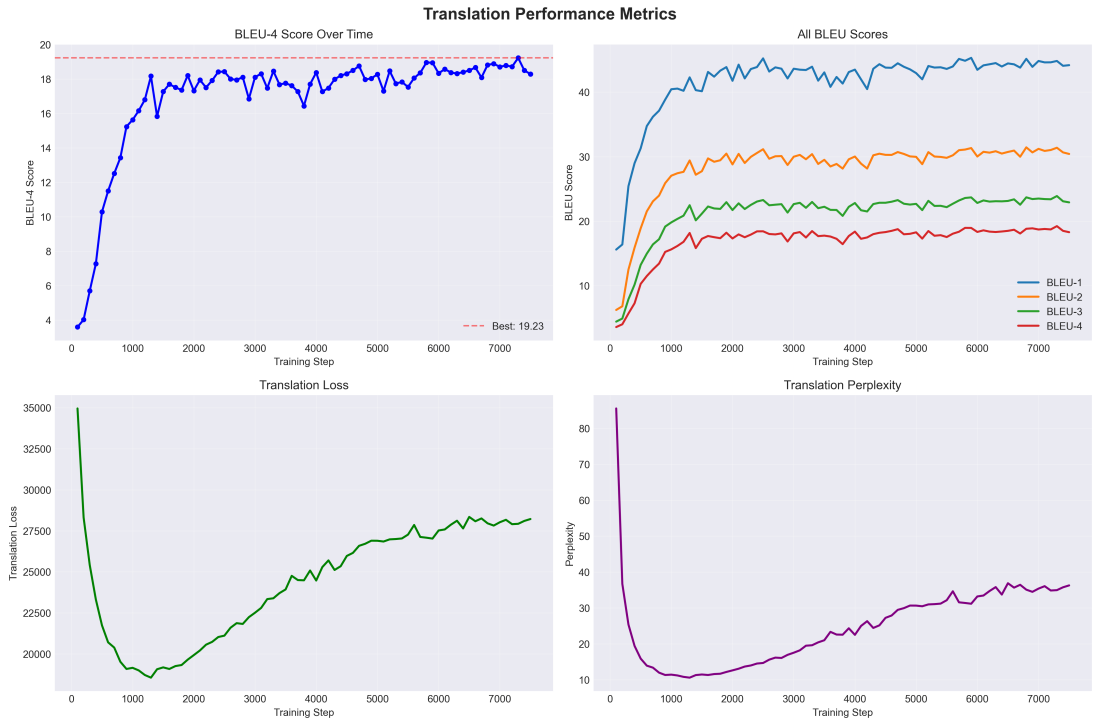
Figure 10

V3 Recongnition Metrics



Figure 11

V3 Translation Metrics



The training behaviour of the V3 model is depicted in Figures 9–11. Similar to V2, V3 utilizes four Transformer layers for both the encoder and decoder components of the model; however, several modifications were implemented in order to stabilize the training process. In comparison with the previous version, the curves for the development set tend to be less volatile, which implies that the training process became more stabile. Additionally, recognition results continue to increase, obtaining lower WER values compared to the early configurations, as well as showing steady growth in translation performance measured via BLEU-4 scores. This observation suggests that the proposed architecture allows the network to learn more consistent visual and linguistic representations without suffering from dramatic drops seen earlier. However, at this point, V3 demonstrates better performance compared to the early versions, but it still does not surpass the performance of the finalized German model, which is confirmed by the achieved results of Model-ODE: WER=31.98%, BLEU-4=21.23. Thus, some further improvements are needed in order to reach the optimal model configuration.

Figure 12
V4 Loss Curves



The behaviour of the training process for Model-V4 is illustrated in Figure 12. This represents the final optimization of the German setup (Model-ODE). Compared to previous designs, Model-V4 utilizes a smaller learning rate, which results in better optimization stability and convergence during training. One important feature of Model-V4 concerns the interrelation between its translation and recognition performance.

As seen from the BLEU versus WER development set plot, the performance of the network improves on both metrics during the late stages of training: the WER drops and BLEU-4 rises. Unlike the translation/recognition trade-offs seen previously, this implies that stronger recognition supervision may improve translation quality instead of adversely affecting it.

As the recognition loss coefficient increases (2.5), the encoder seems to learn how to represent the input for both tasks. Development set WER stabilizes at around 32%, nearly coinciding with the final WER of 31.98% achieved on the test set using optimized beam search decoding. High consistency of the results confirms that

Model-V4 is the optimal design for the German language setting.

Figure 13

V4 Recongnition Metrics



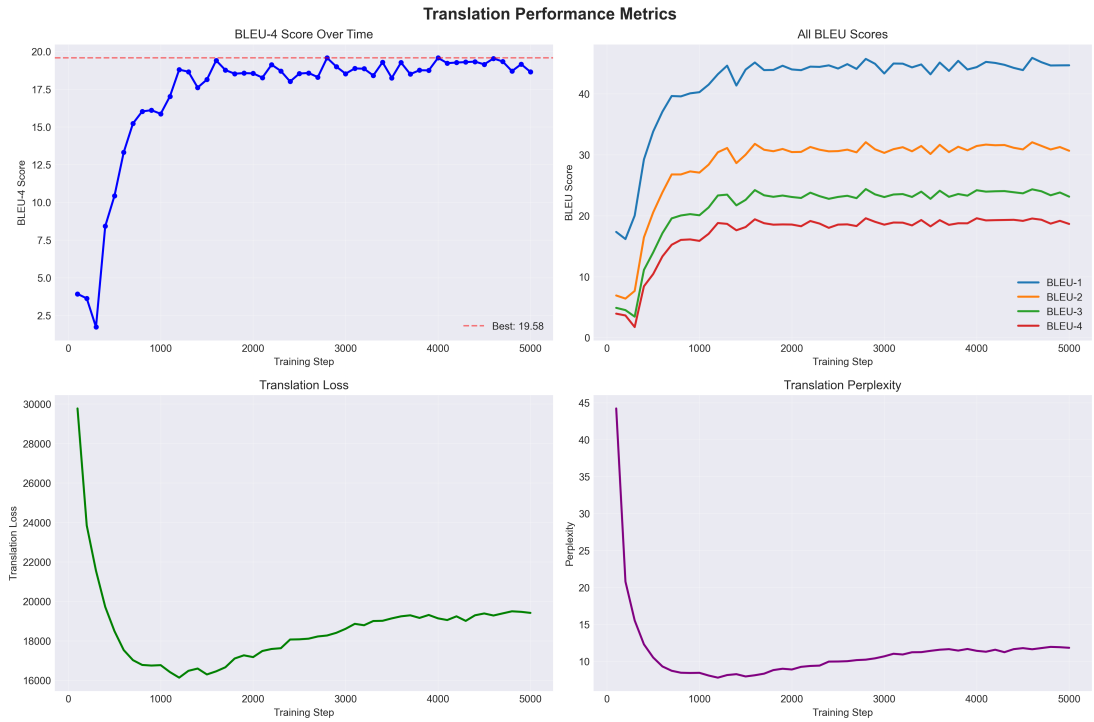
Figure 13 illustrates the improvement in recognition obtained in the final version of V4 (Model-ODE). The most important improvement compared to the previous models is the low level of Word Error Rate (WER), which lies within the low-30% range in development data and equals the final WER on the test set. It can be seen from the error distribution table that the biggest improvement is related to the decrease in the number of deletion errors. The early V1 model often skipped glosses and missed a big share of the reference sequence, while V4 makes many more predictions for the glosses that should appear in the output sequence. Deletions decrease to around 19% on the test set, which is a very significant improvement over previous versions. When deletion rates drop, the number of substitution errors can grow. This pattern was expected since when trying to make many more predictions, some of the predictions will be erroneous, leading to substitutions

However, as the anticipated gloss sequence retains more sign material, this trade-off is typically preferred from a recognition standpoint. The main idea that

recognition tuning is essential to collaborative sign language translation systems is supported by these findings. Stronger overall representations and a more dependable basis for downstream translation are produced by improvements in gloss identification, especially the decrease in deletion errors.

Figure 14

V4 Translation Metrics



Translation performance for the V4 configuration (Model-ODE) is shown in Figure 14. The translation quality is close to the high benchmark obtained by Camgoz et al. (2020) (BLEU-4 score of 21.32). Model-ODE attains a similar BLEU-4 score of 21.23 on PHOENIX-2014-T, suggesting that the proposed architecture achieves competitive spoken language translation quality. While the recognition accuracy is still slightly inferior compared to the system implemented by Camgoz et al. (test WER 31.98% vs. 26.16%), the resulting translation quality remains close. This implies that relatively small discrepancies in gloss recognition quality do not necessarily affect the resulting translation performance, and the translation decoder can produce fluent target language sentences regardless of any residual recognition errors. Moreover, BLEU-4 curves show stable improvement during training. This is expected behavior since label

smoothing is applied to help regularize the translation objective. In summary, the results demonstrate competitive translation performance on the challenging benchmark dataset while using a simpler and fully reproducible training setup

Figure 15

V4-Engl Loss Curves



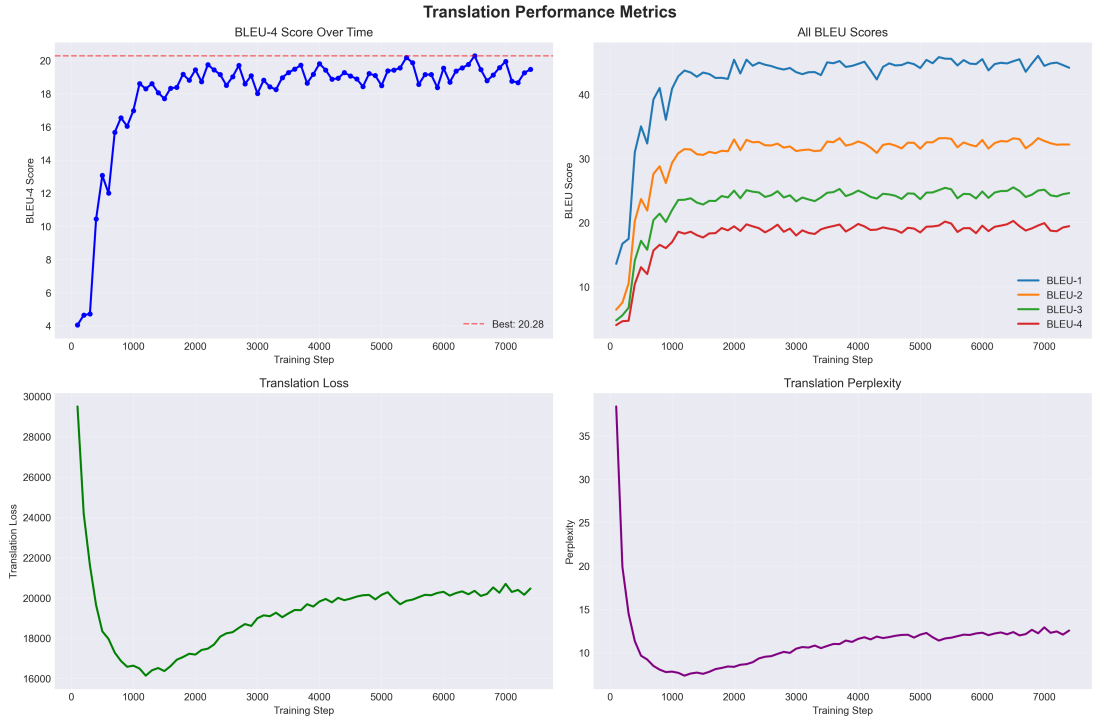
Figure 16

V4-Engl Recongnition Metrics



Figure 17

V4-Engl Translation Metrics



Figures 15-17 show the training behaviour of the V4-English model

(Model-OEN). This model employs the same architecture and hyperparameters as V4 (Model-ODE), yet receives training from the gloss and translation annotations in the PHOENIX release-v3 corpus that pertain to the English dataset. The input visual sign features used by this model are the same as those used in the previous experiments for German recognition. Significantly, the recognition performance in this case is considerably poorer compared to the German experiment.

The final WER on the test set is about 48%, which is significantly higher than the 31.98% of Model-ODE. Yet, the quality of the translation remains about the same, as the test BLEU-4 score is about 21, which is comparable with the one observed on the German model. It should be noted that such observations demonstrate another critical point – high-quality translations can be provided with subpar gloss-recognition quality. Even though there were more gloss-recognition errors on the English side, the sentences created on the basis of this process were still fluent enough to be used for downstream tasks. This example is crucial for the multilingual cascade approach mentioned in Section 4.2, where English outputs of Model-OEN act as usable pivot sentences for Stage 1. Figures 15 to 17 show that it is possible to develop an English text generator that could be used in combination with multilingual systems like NLLB.

4.1.2 Prior Work Baseline (German Test Set)

For direct comparison to the CVPR 2020 backbone on PHOENIX-2014-T (German):

Table 3
Baseline Results on PHOENIX-2014-T (German Test Set).

Reference model (Camgoz et al., 2020)	Test WER ↓	Test BLEU-4 ↑
Sign2(Gloss+Text), joint	26.16	21.32
Sign2Text (direct)	45.34	20.17

4.1.3 Unified Multilingual DE+EN Results (Shared Model)

We train a single unified DE+EN model with language tags (`<de>`, `<en>`) on aligned bilingual manifests (`PHOENIX2014T_DE_EN_V2`). Metrics uses the same

BLEU/WER tooling as Sections 6.1–6.3.

Model progression: DE+EN-v2 is trained from scratch, DE+EN-v3 is fine-tuned from v2, and DE+EN-v4 (lang-head) is fine-tuned from v3 with language-specific translation output heads.

Model selection: the training checkpoint achieving the highest validation BLEU-4 under greedy decoding (beam 1).

Decode tuning: CTC beam size and translation beam size/α are grid-searched on the development set and applied once to the test set without access to test labels.

Table 4
Unified Multilingual DE+EN Model Results.

Model	Dev WER ↓	Test WER ↓	Dev BLEU-4 ↑	Test BLEU-4 ↑	Best decoding setup
DE+EN-v2 (single softmax)	73.98	73.24	12.12	11.87	CTC BW=9, TX BW=4, α=3
DE+EN-v3 (single softmax)	68.35	68.08	12.88	12.21	CTC BW=10, TX BW=6, α=3
DE+EN-v4 (lang-head, f.t. from v3)	66.97	66.04	24.23	24.67	CTC BW=10, TX BW=8, α=3

For DE+EN-v4, the best checkpoint according to the training criterion was obtained at optimisation step 13,600, achieving a validation BLEU-4 score of 23.18 under greedy decoding. Decode-grid refinement on the development set further improved the dev BLEU-4 score to **24.23**.

Using the same tuned decoding configuration on the test set, the model achieved a **WER of 66.04** (DEL 26.91, INS 1.18, SUB 37.96), **BLEU-4 of 24.67**, CHRF of 47.98, and ROUGE of 50.70.

Table 5
Comparison of Monolingual and Multilingual SLT Models.

Model	Dev WER	Test WER	Dev BLEU-4	Test BLEU-4	Decode / notes
Camgoz et al. (2020) Sign2(Gloss+Text), joint	–	26.16	–	21.32	German test (published baseline)
Model-O_{DE} (optimised DE)	32.03	31.98	20.62	21.23	German supervision; dev-selected decoding
Model-O_{EN} (optimised EN)	47.52	48.39	21.09	20.88	release-v3 EN labels; CTC BW=8, TX BW=4, α=2
DE+EN-v2 (single softmax)	73.98	73.24	12.12	11.87	CTC BW=9, TX BW=4, α=3
DE+EN-v3 (single softmax; f.t. from v2)	68.35	68.08	12.88	12.21	CTC BW=10, TX BW=6, α=3
DE+EN-v4 (lang-head; f.t. from v3)	66.97	66.04	**24.23**	**24.67**	CTC BW=10, TX BW=8, α=3; best ckpt at step 13,600

Table 6
Per-language DE+EN-v4 Metrics on The PHOENIX-2014-T Test Set.

Subset	Test WER ↓	Test BLEU-4 ↑	n
All (bilingual)	66.04	24.67*	1284
<de> only	59.69	22.54	642
<en> only	77.14	17.90	642

Note: Corpus BLEU computed over the combined bilingual test set containing both German and English references.

Per-language evaluation.

Table 6 reports test-set performance for the DE+EN-v4 model separated by output language. The German subset achieves a BLEU-4 score of 22.54, which is close to the monolingual German Model-ODE result of 21.23.

On the other hand, the English subset yields BLEU-4 score of 17.90, still lower than the monolingual English Model-OEN score of 20.88. The recognition capability is much poorer compared to the monolingual models, with WER scores of 59.69% and 77.14%, respectively, for the German and English languages. The poor recognition scores can be attributed to the usage of a common CTC gloss predictor in both languages. Since the recognition part is not language specific, it causes

competition among gloss representation for both the languages in the output space, resulting in errors in recognition and language interference. Despite high WER scores, translation performance stays fairly high, especially for the German subset. This indicates that the multilingual model has learned valuable translation representations even with reduced gloss recognition performance for both the languages, but English translations are harder than German in bilingual case.

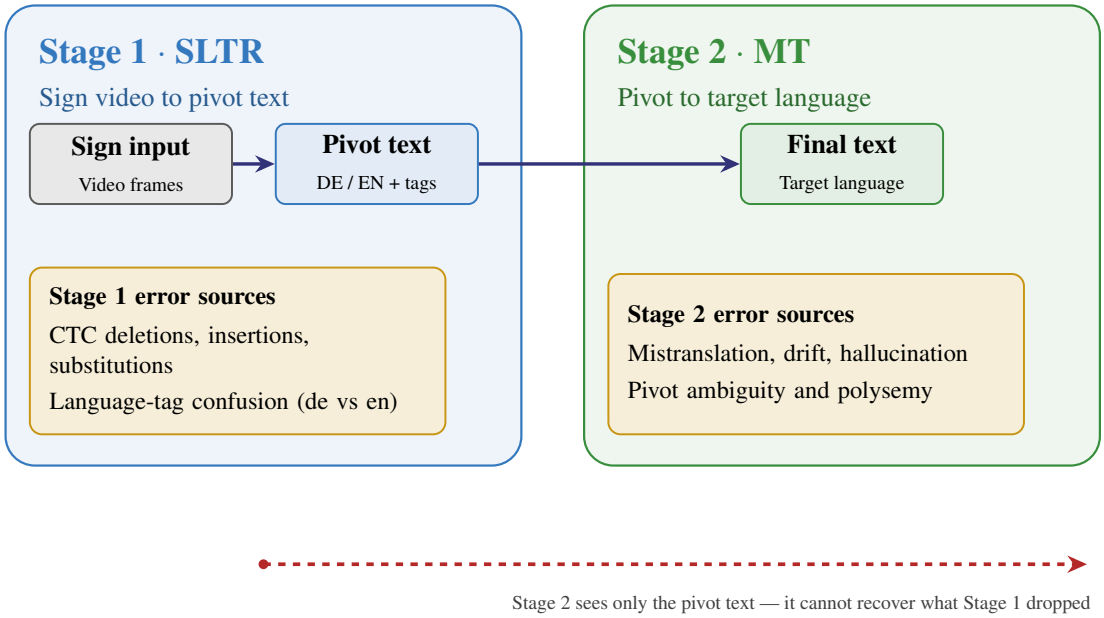
4.1.4 Discussion Of The Current Bottleneck

DE+EN-v2 and **DE+EN-v3** unified multilingual models were found to have considerable cross-language interference when using a single softmax layer shared by both languages despite the presence of language tags. The bilingual dev/test BLEU scores were lower than the baseline monolingual results **(Model-ODE and Model-OEN)**, showing that using the same output distribution prevented language-specific generation. In an attempt to solve the problem, **DE+EN-v4** uses target language-specific output translation heads while retaining the encoder-decoder Transformer architecture that is completely shared among languages but only the final translation head per target language is specific to each individual language.

It is worth noting that the BLEU-4 performance measure of 24.67 reported for the DE+EN-v4 configuration on the bilingual evaluation dataset consisting of German and English references. Accordingly, the obtained results cannot be directly compared to the German only PHOENIX-2014-T evaluation datasets used in previous studies (e.g. CVPR 2020 baseline). The much higher value of WER in the unified setup (66.04 compared to 26.16) is caused by the higher difficulty of bilingual recognition and joint-vocabulary optimisation. Despite that, compared to the unified models of DE+EN-v2 and DE+EN-v3, the new DE+EN-v4 model demonstrates significantly better BLEU performance together with consistent decreases in WER due to using language-conditioned output heads without additional visual networks

Figure 18

Error Propagation in The Two-stage Cascaded Sign-to-target Translation Pipeline.



The sources of errors in the proposed two-stage cascaded architecture are shown in Figure 18. The quality of translation depends on the errors that can happen in the entire process. The first error source is sign recognition. Errors in recognition, indicated by large WER values, could result in glosses being incorrectly, partially, or completely missed. Information lost at this step cannot be salvaged during the subsequent stages of the pipeline. The second error source is the pivot text generation. In the case of correct sign recognition, the errors introduced during stage one could include repetitions, truncations, language tag mistakes, or poor English language generation in the bilingual scenario. The errors affect the quality of the information transferred to the subsequent step.

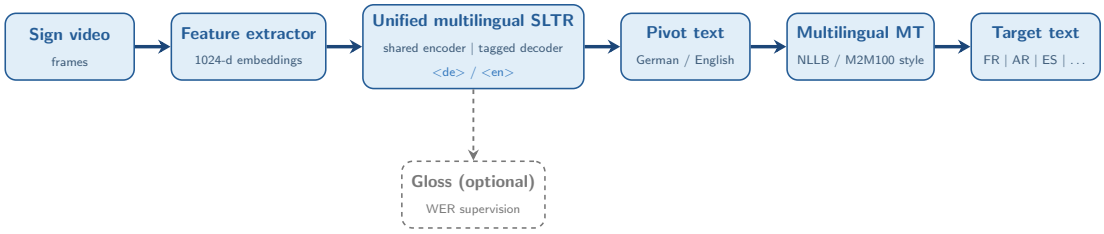
The multilingual translation approach may itself have problems like mistranslation, lexical drift or hallucinations. However, Stage 2 can work only on the pivot text and does not have any access to the original signing video and visual signals. Therefore, while Stage 2 can often help fix minor fluency problems or grammatical issues within the pivot sentence, it cannot compensate for information that was lost prior to the input into Stage 2. For instance, if the English pivot sentence had ended in an unfinished clause like "…and tomorrow" without the end of the

phrase, then the French output would still end in an unfinished clause because the necessary information was not present in the Stage-2 input.

4.2 Cross-Lingual Cascade Extension

The suggested cascade system extends the capability of sign language translation to languages other than German and English that the DE+EN-v4 system can directly translate into. As shown in Figure 18, the sign language translation system includes the following two steps:

Figure 19
End-to-end Multilingual Cascade Framework. An Unified DE+EN SLTR model.



The first stage involves generation of a pivot phrase in German or English using the DE+EN-v4 sign language translation system. In the second step, the pivot text is translated into the target language using a multilingual machine translation system. The multilingual translation system works only on the text and has no information about the original sign video input. Thus, there can be cascaded errors as indicated in Figure 18.

4.2.1 Protocol

Pivot language.

English serves as the default pivot language since it is the source language for the multilingual machine translation systems like NLLB. German pivots are still possible using the ¡de¿ language tag and produce better results at stage one translation on the PHOENIX-2014-T (BLEU-4 22.54 vs. 17.90 using English pivot 6).

Stage-2 model. uses the NLLB-200-distilled-600M model“Scaling neural machine translation to 200 languages” (2024) for multilingual text translation.

Input preparation. By choosing samples with the en suffix and eliminating the <en> language tag, English pivot sentences are retrieved from the DE+EN-v4 test-set hypothesis file.

Evaluation. For French and other cascade target languages, PHOENIX-2014-T does not offer reference translations. As a result, evaluation is solely qualitative, and Stage 2 outputs do not have BLEU scores.

4.2.2 Pilot Translation Results (EN → FR)

Table 7 presents five example outputs from the English-to-French cascade.

Table 7
Pilot Cascade Results: English Pivots Generated By DE+EN-v4 and Translated to French Using NLLB.

#	Stage-1 pivot (EN)	Stage-2 output (FR)
1	the rain clouds will at the edge of the alps during the day in the northeast and the stars will appear here and there.	Les nuages de pluie seront au bord des Alpes pendant la journée dans le nord-est et les étoiles apparaîtront ici et là.
2	on thursday it will rain in the northeast, otherwise there will be sun, clouds and on friday it will be rain, especially on friday	Jeudi, il pleuvra dans le nord-est, sinon il y aura du soleil, des nuages et vendredi, il pleuvra, surtout vendredi.
3	a new low pressure system will bring us heavy snowfall and tomorrow	Un nouveau système de basse pression nous apportera de fortes chutes de neige et demain
4	in the morning it will be clear and up to seventeen degrees in the lower rhine	Le matin, il sera clair et jusqu'à dix-sept degrés dans le bas du Rhin.
5	germany will be in the influence of high pressure and clouds will cross us tomorrow	L'Allemagne sera sous l'influence de la pression et les nuages nous traverseront demain

Good English pivots (rows 1, 2, and 4) usually lead to good French translations. However, any mistake in the pivot sentence is immediately carried over to the output. In the case of row 3, where the pivot sentence is incomplete in English, the French translation is incomplete as well. Likewise, the meteorological phraseology in row 5, though grammatically correct in English, is translated imprecisely in French.

4.2.3 Discussion

The cascade architecture enables multilingual sign language translation without retraining the visual sign-language model. Once the bilingual DE+EN-v4 model has

been trained, additional target languages can be supported by replacing or extending the Stage 2 machine translation component.

The present pilot study has several limitations. First, the analysis is based on only five example sentences and is not intended to provide statistically representative results. Second, no quantitative Stage 2 evaluation is possible because PHOENIX does not provide French reference translations. Third, the NLLB model is used without fine-tuning on sign-derived pivot text and therefore may be sensitive to recognition and translation noise introduced by Stage 1.

CHAPTER 5
CONCLUSION AND IMPLICATIONS

5.1 Conclusion

To support multilingual experimentation, we reconstructed English supervision from PHOENIX-2014-T release-v3 with complete split alignment and trained an English-supervised model (**Model-OEN**). Unified DE+EN training under a single translation softmax (DE+EN-v2/v3) revealed persistent cross-language interference despite language tagging. The DE+EN-v4 variant, which introduces language-specific translation heads while preserving a shared encoder, shared decoder trunk, and a single CTC gloss branch, produced a large improvement in bilingual translation quality, reaching a bilingual test BLEU-4 score of **24.67**. Recognition performance also improved, although unified bilingual WER remained high (**66.04**), reflecting the increased difficulty of multilingual gloss recognition under shared supervision. We therefore emphasise that bilingual scores should not be interpreted as directly comparable to German-only published benchmarks.

Finally, We implemented and piloted the proposed two-stage cascade architecture (Section 4.2), translating five English pivot sentences generated by DE+EN-v4 into French using NLLB. The pilot demonstrates that downstream translation quality is strongly dependent on the completeness and accuracy of the Stage 1 pivot text. Fluent and well-formed pivots generally produce coherent target-language outputs, whereas truncation and semantic errors propagate directly into the final translation.

The results also highlight a key advantage of the cascade approach: additional written languages can be supported through the multilingual machine translation stage without retraining the visual sign-language encoder or collecting new sign-language annotations.

5.2 Future Work

1. Scale the proposed cascade to the full English test split and extend evaluation to additional target languages supported by NLLB-200.

2. Introduce a language-conditioned CTC recognition head to better separate German and English gloss prediction and reduce cross-language interference in the recognition branch.

3. Systematically compare gloss-based versus pivot-text-based inputs for Stage 2 machine translation, in order to quantify the impact of representation choice on downstream translation quality.

4. To provide a more thorough review of cascade translation quality than BLEU-based measurements, include sophisticated evaluation metrics like COMET.

5.3 Data and Code Availability

PHOENIX-2014-T is available under an open license. Training setups, preprocessing, and experimental code have been included in the project repository for reproducibility. Machine translation experiments in Stage 2 utilize the facebook/nllb-200-distilled-600M model "Scaling neural machine translation to 200 languages" (2024), provided by the Hugging Face Transformers library

REFERENCES

Attar, R. K., Goyal, V., & Goyal, L. (2022). Open issues and challenges in automatic sign language translation systems. *AIP Conference Proceedings*, *2455*(1), 040003.

Babour, A., Bitar, H., Alzamzami, O., Alahmadi, D., Barsheed, A., Alghamdi, A., & Almshjary, H. (2023). Intelligent gloves: An it intervention for deaf-mute people. *Scientific Reports*, *13*(1), 1–15.

Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., & Bowden, R. (2018). Neural sign language translation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7784–7793.

Camgoz, N. C., Koller, O., Hadfield, S., & Bowden, R. (2020). Sign language transformers: Joint end-to-end sign language recognition and translation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10023–10033.

Chen, Y., Wei, F., Sun, X., Wu, Z., & Lin, S. (2022). A simple multi-modality transfer learning baseline for sign language translation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5120–5130.

Harshitha, C., Rajagopalan, N., et al. (2024). A survey on machine and deep learning approaches in sign language recognition: Techniques and future trends. *2024 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT)*, 1–6.

Hu, L., Gao, L., Liu, Z., & Feng, W. (2023). Continuous sign language recognition with correlation network. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2529–2539.

Johnson, M., Schuster, M., Le, Q. V., et al. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the ACL*, *5*, 339–351.

Kermani, A., Irani, H., & Metsis, V. (2025). Finetuning pre-trained language models for bidirectional sign language gloss to text translation. *Proceedings of the Workshop on Sign Language Processing (WSLP)*, 73–81.

Li, D., Xu, C., Yu, X., Zhang, K., Swift, B., Suominen, H., & Li, H. (2020). Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. *Advances in Neural Information Processing Systems*, *33*, 12034–12045.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, *8*, 726–742.

Mohsin, S., Salim, B. W., Mohamedsaeed, A. K., Ibrahim, B. F., & Zeebaree, S. R. (2023). American sign language recognition based on transfer learning algorithms. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(5s), 390–399.

Nandi, U., Ghorai, A., Singh, M. M., Changdar, C., Bhakta, S., & Kumar Pal, R. (2023). Indian sign language alphabet recognition system using cnn with diffgrad optimizer and stochastic pooling. *Multimedia Tools and Applications*, *82*(7), 9627–9648.

Obi, Y., Claudio, K. S., Budiman, V. M., Achmad, S., & Kurniawan, A. (2023). Sign language recognition system for communicating to people with disabilities. *Procedia Computer Science*, *216*, 13–20.

Papatsimouli, M., Sarigiannidis, P., & Fragulis, G. F. (2023). A survey of advancements in real-time sign language translators: Integration with iot technology. *Technologies*, *11*(4), 83.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Scaling neural machine translation to 200 languages. (2024). *Nature*, *630*(8018), 841–846.

Singh, S. K., & Chaturvedi, A. (2023). A reliable and efficient machine learning pipeline for american sign language gesture recognition using emg sensors. *Multimedia Tools and Applications*, *82*(15), 23833–23871.

Sperber, M., & Paulik, M. (2020). Speech translation and the end-to-end promise: Taking stock of where we are. *Proceedings of ACL*, 7409–7421.

Sreemathy, R., Turuk, M., Jagdale, J., Agarwal, A., & Kumar, V. (2023). Indian sign language interpretation using convolutional neural networks. *2023 10th International Conference on Signal Processing and Integrated Networks (SPIN)*, 789–794.

Wang, A., Li, J., Luan, W., & Pan, L. (2025). OVSLT: Advancing sign language translation with open vocabulary. *Electronics*, *14*(5). https://doi.org/10.3390/electronics14051044

World Health Organisation. (2024). *Deafness and hearing loss* (tech. rep.). World Health Organisation. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). Mt5: A massively multilingual pre-trained text-to-text transformer. *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*, 483–498.

Zhou, B., Chen, Z., Clapés, A., Wan, J., Liang, Y., Escalera, S., Lei, Z., & Zhang, D. (2023). Gloss-free sign language translation: Improving from visual-language pretraining. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20871–20881.

APPENDICES

APPENDIX A: SIMILARITY REPORT

7% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Small Matches (less than 8 words)

Exclusions

- 3 Excluded Sources

Match Groups

48 Not Cited or Quoted 5%
Matches with neither in-text citation nor quotation marks

15 Missing Quotations 2%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

5% Internet sources
4% Publications
1% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **48** Not Cited or Quoted 5%
  Matches with neither in-text citation nor quotation marks
- **15** Missing Quotations 2%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

5% Internet sources
4% Publications
1% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	aclanthology.org <1%
2	Internet	docs.neu.edu.tr <1%
3	Internet	arxiv.org <1%
4	Internet	www.mdpi.com <1%
5	Internet	dspace.ut.ee <1%
6	Internet	files.sciconf.cn <1%
7	Internet	www.statmt.org <1%
8	Internet	www2.statmt.org <1%
9	Student papers	King's College <1%
10	Publication	Kusubakti Andajani, Sari Karmina, Lidya Amalia Rahmania. "Inclusive, Sustainabl... <1%

11 Student papers
University of Stirling <1%

12 Student papers
University of Sydney <1%

13 Student papers
Bright Business School <1%

14 Student papers
UC, Irvine <1%

15 Publication
"Preface", IOP Conference Series: Materials Science and Engineering, 2021 <1%

16 Student papers
RMIT University <1%

17 Internet
eprints.qut.edu.au <1%

18 Publication
Harshitha C, Sendil Vadivu D, Narendran Rajagopalan. "A Survey on Machine and ... <1%

19 Publication
Mukushev, Medet. "Towards Large-Vocabulary Kazakh Sign Language Processing:... <1%

20 Publication
Zuo, Ronglai. "Vision-Based Sign Language Processing : Recognition, Translation, ... <1%

21 Internet
dspace.lib.ntua.gr <1%

22 Internet
ojs.aaai.org <1%

23 Internet
ufal.mff.cuni.cz <1%

24 Internet
wesscholar.wesleyan.edu <1%

25	Publication	Batuhan Baykara, Tunga Güngör. "Turkish abstractive text summarization using ... <1%
26	Publication	Büşra İçöz, Göksel Biricik. "Context-Aware Code Review Automation: A Retrieval-... <1%
27	Publication	Jungeun Kim, Ha Young Kim. "CSLT-AK: Convolutional-embedded Transformer wit... <1%
28	Internet	lrec-conf.org <1%
29	Internet	umt-ir.umt.edu.my:8080 <1%
30	Internet	vbn.aau.dk <1%
31	Publication	"Proceedings of World Conference on Artificial Intelligence: Advances and Applic... <1%
32	Publication	Changtong Zan, Liang Ding, Li Shen, Yibin Lei, Yibing Zhan, Weifeng Liu. "Improvi... <1%
33	Publication	Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, Richard Bowden. "Sign Langu... <1%
34	Publication	Paolo Ferro, Harinadh Vemanaboina, Chander Prakash. "Computational Techniqu... <1%
35	Internet	dadelani.github.io <1%
36	Internet	eprints.utm.my <1%
37	Internet	mdpi-res.com <1%
38	Internet	os.unil.cloud.switch.ch <1%

39 Internet

tdx.cat <1%

40 Internet

www.medrxiv.org <1%

41 Internet

www2.mdpi.com <1%

AI-Powered Assistive Translator Designed For Deaf-Mute Individuals

Imane Boumedra

Final International University
June 2026
Girne, TRNC

Imane Boumedra
AI-Powered Assistive Translator Designed For Deaf-Mute Individuals
MS 2026

APPENDIX B: STATEMENT OF ARTIFICIAL INTELLIGENCE USAGE

To the Institute of Graduate Education of Final International University,

I hereby declare that the subject matter, methodology, analyses, interpretations, and conclusions of this research are entirely my own work. I accept full academic, ethical, and legal responsibility for this study. Furthermore, I declare and undertake that I have not used artificial intelligence at any stage of the aforementioned processes.

Sincerely,

Full Name: **IMANE BOUMEDRA**

Student ID Number: **2410140005**

Signature:

Date: **02/07/2026**