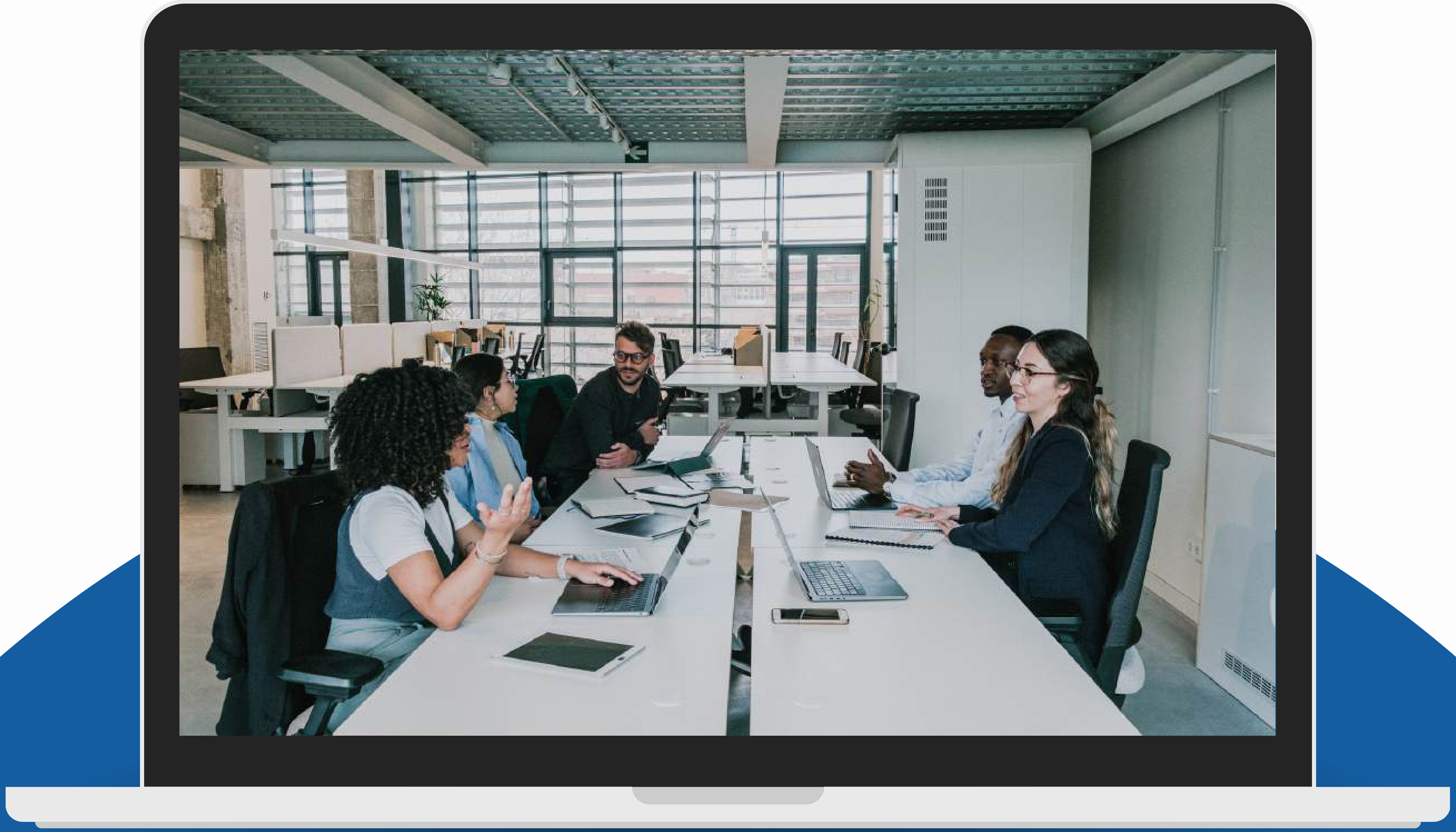
# Sign Language Translation System using Machine Learning



FINAL REPORT

## AI PROJECT

Group 09

# Members

| | |
|---|---|
| Tạ Khánh Hà | 31221023776 |
| Nguyễn Xuân Hân | 31221023021 |
| Nguyễn Phương Nghi | 31221020433 |
| Võ Trần Bảo Trân | 31221024798 |
| Nguyễn Thị Cẩm Tú | 31221021819 |

# Overview

# 1. Introduction

## 1.1. Project Objectives

Develop a multilingual **sign language translation tool**, akin to **Google Translate**, that supports three primary languages: **English**, **German**, and **French**. Users will be able to input text, and the tool will generate a corresponding sign language video. The tool will also allow translation from sign language to text by uploading videos.

## 1.2. Rationale for Selecting the Topic

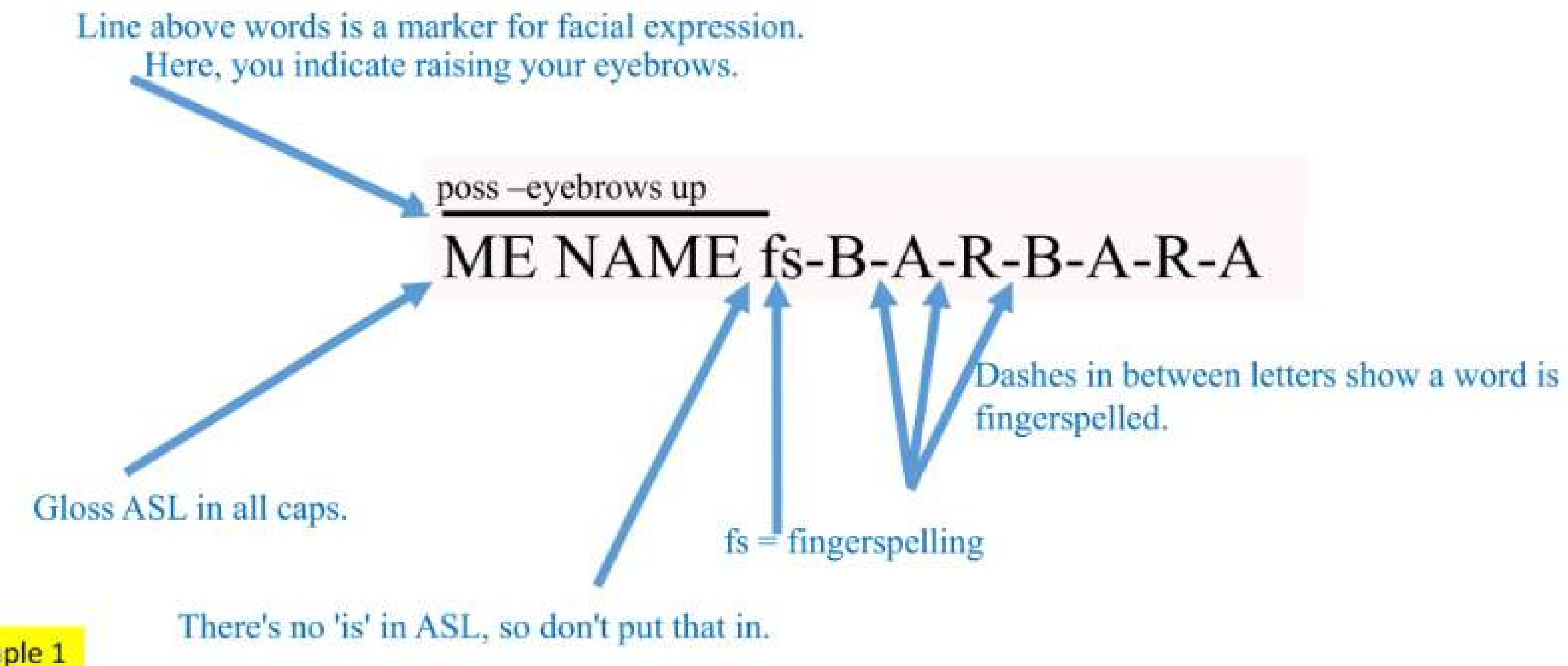
- While automated translation technologies like Google Translate help overcome language barriers, **sign language remains overlooked**.
- The lack of translation tools poses significant communication challenges for the hearing impaired and hinders those wanting to learn sign language due to **limited resources**.
- The project will enhance education and research, contributing to the establishment of a society free from language barriers and promoting linguistic diversity and equality.

# 2. Literature Review & Related works

**2.1. Gloss-Based Sign Language Translation Models**

**2.1. Gloss-Free Sign Language Translation Models**

# What does ASL **glossing** look like?




Example 1

## 2.1. Gloss-Based Sign Language Translation Models

| Model name | Description | Reference |
|---|---|---|
| 1. Neural Sign Language Translation | • Camgoz et al., 2018 introduced a model using **CNNs** and **Bi-LSTM networks** for Sign2Gloss and an **attention-based Seq2Seq** model for Gloss2Text.<br>• **Strengths:** Structured learning; improved accuracy.<br>• **Weaknesses:** Annotation dependence; error propagation. | Camgoz, N. C., Koller, O., Hadfield, S., & Bowden, R. (2018). Neural Sign Language Translation. CVPR, 7784-7793. |
| 2. Sign Language Transformer | • Yin et al., 2020 introduced a transformer-based model leveraging gloss annotations: Transformer Architecture and Multi-Task Learning.<br>• **Strengths:** Enhanced contextual understanding; joint optimization.<br>• **Weaknesses:** High computational; data demands. | Yin, S., Xia, Z., Chen, X., Zhou, H., & He, S. (2020). Sign Language Translation with Transformer. MM '20, 1778–1786. |
| 3. Neural Sign Language Translation by Learning Tokenization | • Orbay & Akarun, 2020 introduced a model that learns sub-word tokenization of glosses, with Tokenization Mechanism and End-to-End Training.<br>• **Strengths:** Vocabulary efficiency; flexibility.<br>• **Weaknesses:** Added complexity; dependence on gloss annotation quality. | Orbay, E., & Akarun, L. (2020). Neural Sign Language Translation by Learning Tokenization. arXiv:2004.03519. |

## 2.2. Gloss-Free Sign Language Translation Models

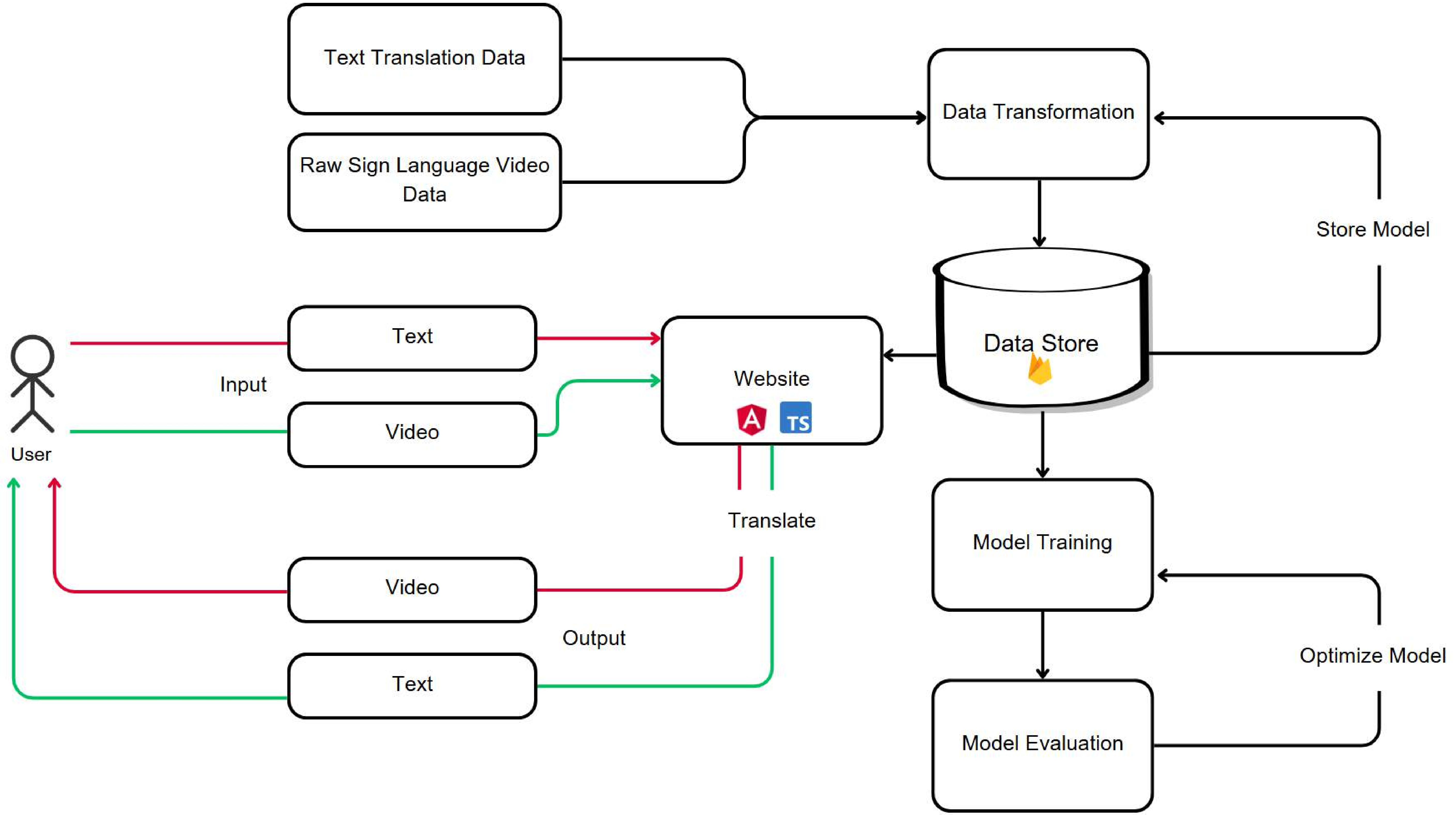
| Model name | Description | Reference |
| --- | --- | --- |
| 1. Sign Language Transformers | • Camgoz et al., 2020 developed a gloss-free transformer model with **End-to-End Architecture** and **Spatial-Temporal Encoder** plus a **Transformer Decoder**.<br>• **Strengths:** No gloss annotations needed; nuanced capture of sign input.<br>• **Weaknesses:** Data intensiveness; training complexity. | Camgoz, N. C., Hadfield, S., Koller, O., & Bowden, R. (2020). Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation. CVPR, 10023-10033. |
| 2. Progressive Transformers | • Saunders et al., 2020 proposed a progressive transformer model with **Progressive Learning** and **Multi-Cue Integration**.<br>• **Strengths:** Enhanced feature learning; rich contextual information.<br>• **Weaknesses:** Increased model complexity; high resource demand. | Saunders, B., Camgoz, N. C., & Bowden, R. (2020). Progressive Transformers for End-to-End Sign Language Production. ECCV, 687–705. |
| 3. Self-Supervised Learning for SLT | • Pu et al., 2019 explored self-supervised learning for gloss-free SLT using **Cross-Modal Training** and **Feature Extraction**.<br>• **Strengths:** Data efficiency; improved generalization.<br>• **Weaknesses:** Limited improvement without large datasets; additional design complexity. | Pu, J., Zhou, P., Wang, F., & Xu, W. (2019). Boosting Continuous Sign Language Recognition via Cross Modality Augmentation. CVPR, 11509-11518. |

# 3. Methology

3.1 System Architecture Overview

3.2 AI model Architecture Overview

# 3.1 System Architecture Overview

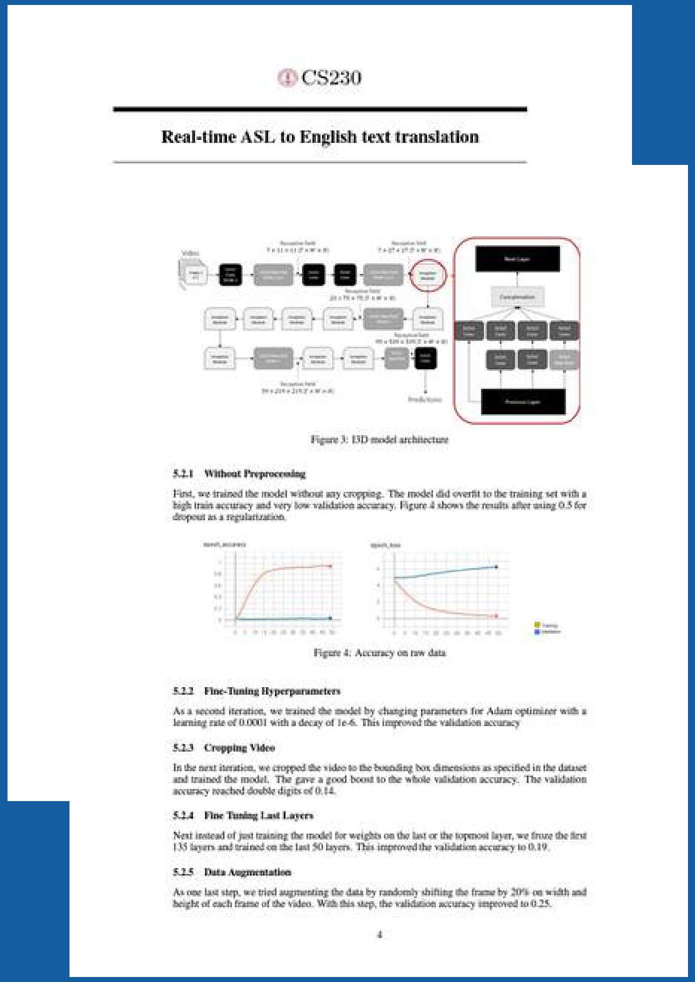


## 3.2 AI model Architecture Overview

# Approach 1

MS-ASL Dataset

+



CS230

**Real-time ASL to English text translation**

Figure 3: I3D model architecture

**5.2.1 Without Preprocessing**

First, we trained the model without any cropping. The model did overfit to the training set with a high train accuracy and very low validation accuracy. Figure 4 shows the results after using 0.5 for dropout as a regularization.

Figure 4: Accuracy on raw data

**5.2.2 Fine-Tuning Hyperparameters**

As a second iteration, we trained the model by changing parameters for Adam optimizer with a learning rate of 0.0001 with a decay of 1e-6. This improved the validation accuracy

**5.2.3 Cropping Video**

In the next iteration, we cropped the video to the bounding box dimensions as specified in the dataset and trained the model. The gave a good boost to the whole validation accuracy. The validation accuracy reached double digits of 0.14.

**5.2.4 Fine Tuning Last Layers**

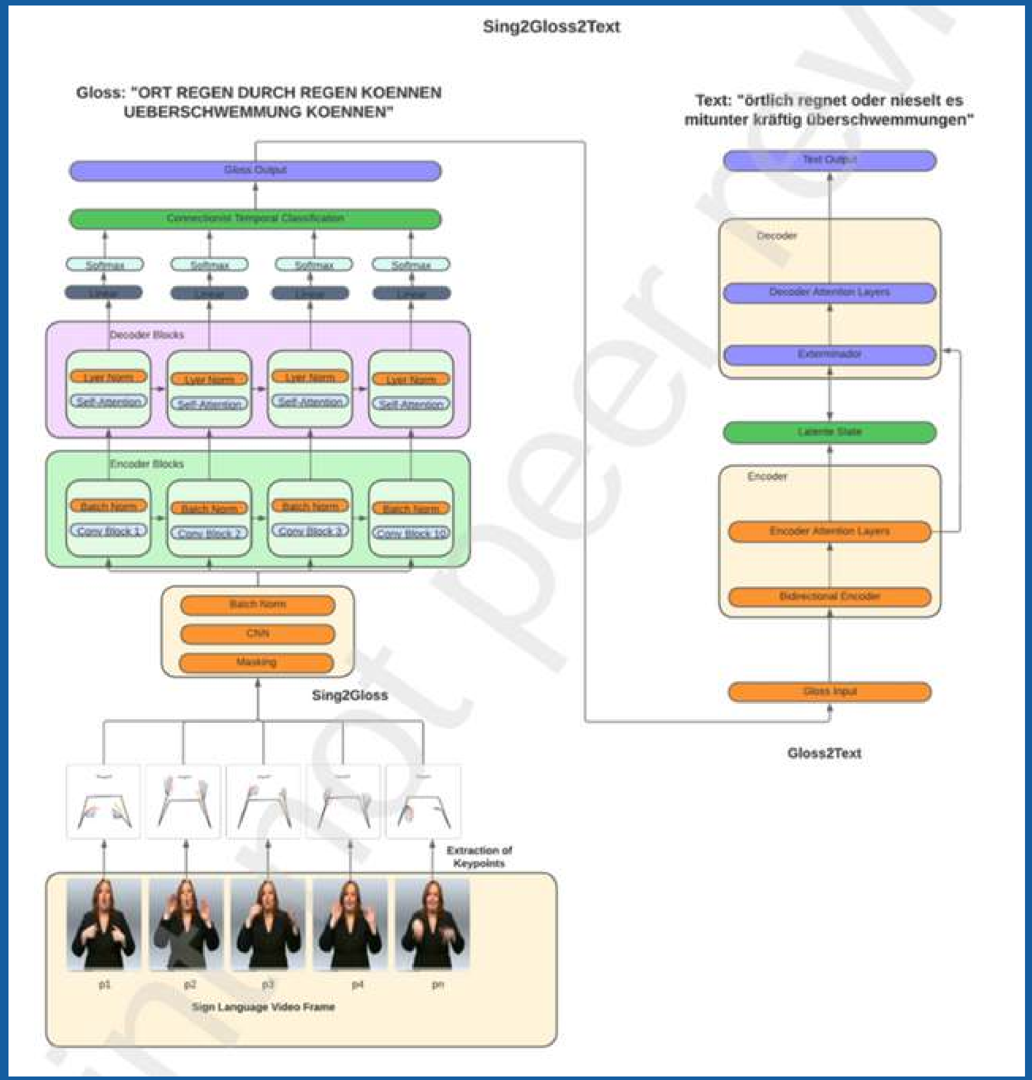
Next instead of just training the model for weights on the last or the topmost layer, we froze the first 135 layers and trained on the last 50 layers. This improved the validation accuracy to 0.19.

**5.2.5 Data Augmentation**

As one last step, we tried augmenting the data by randomly shifting the frame by 20% on width and height of each frame of the video. With this step, the validation accuracy improved to 0.25.

4

I3D (Standford)




Sign2Gloss2Text

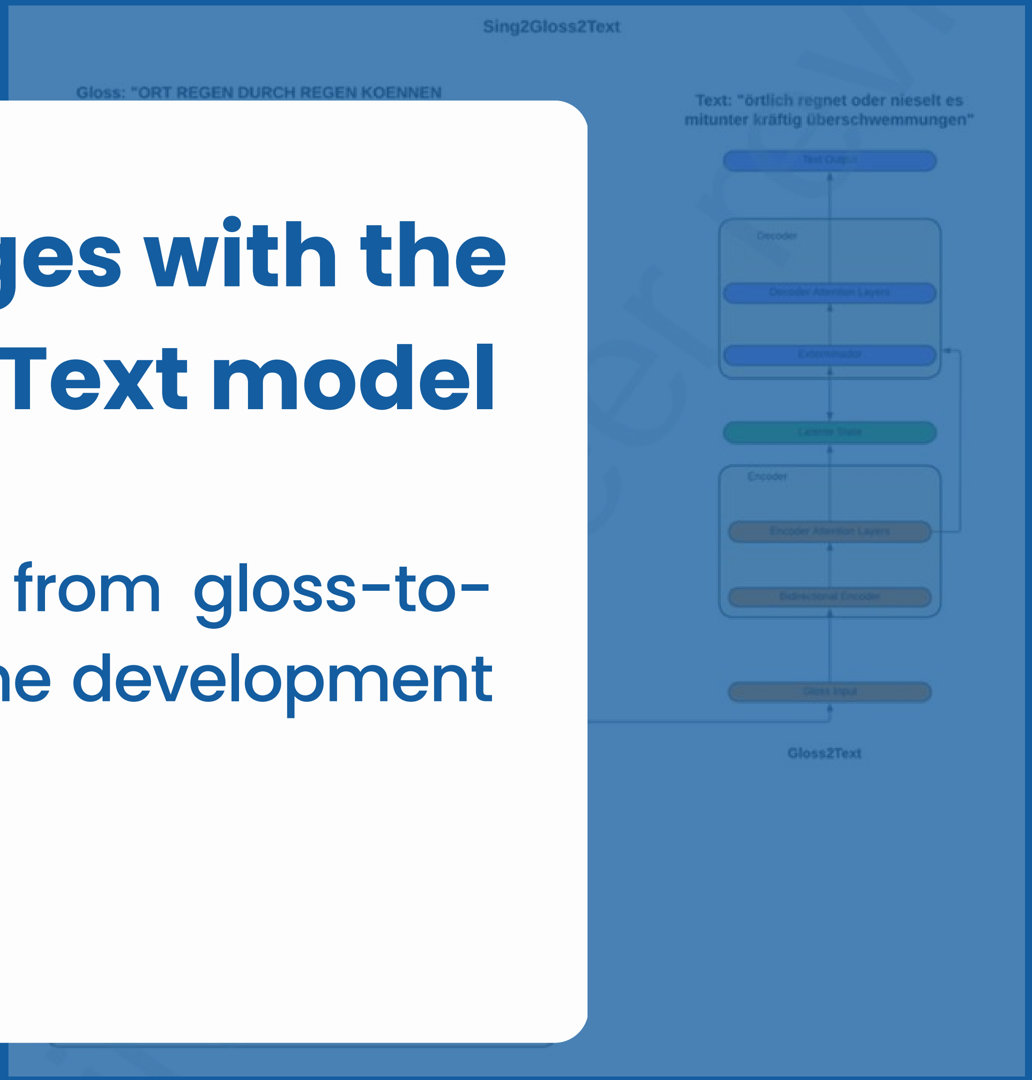
# Approach 1

## The problem with the MS-ASL dataset

Only about 1,600 out of more than 16,000 vocabulary items were successfully downloaded.

## The challenges with the Sign2Gloss2Text model

The lack of data from gloss-to-text has slowed the development process of model.



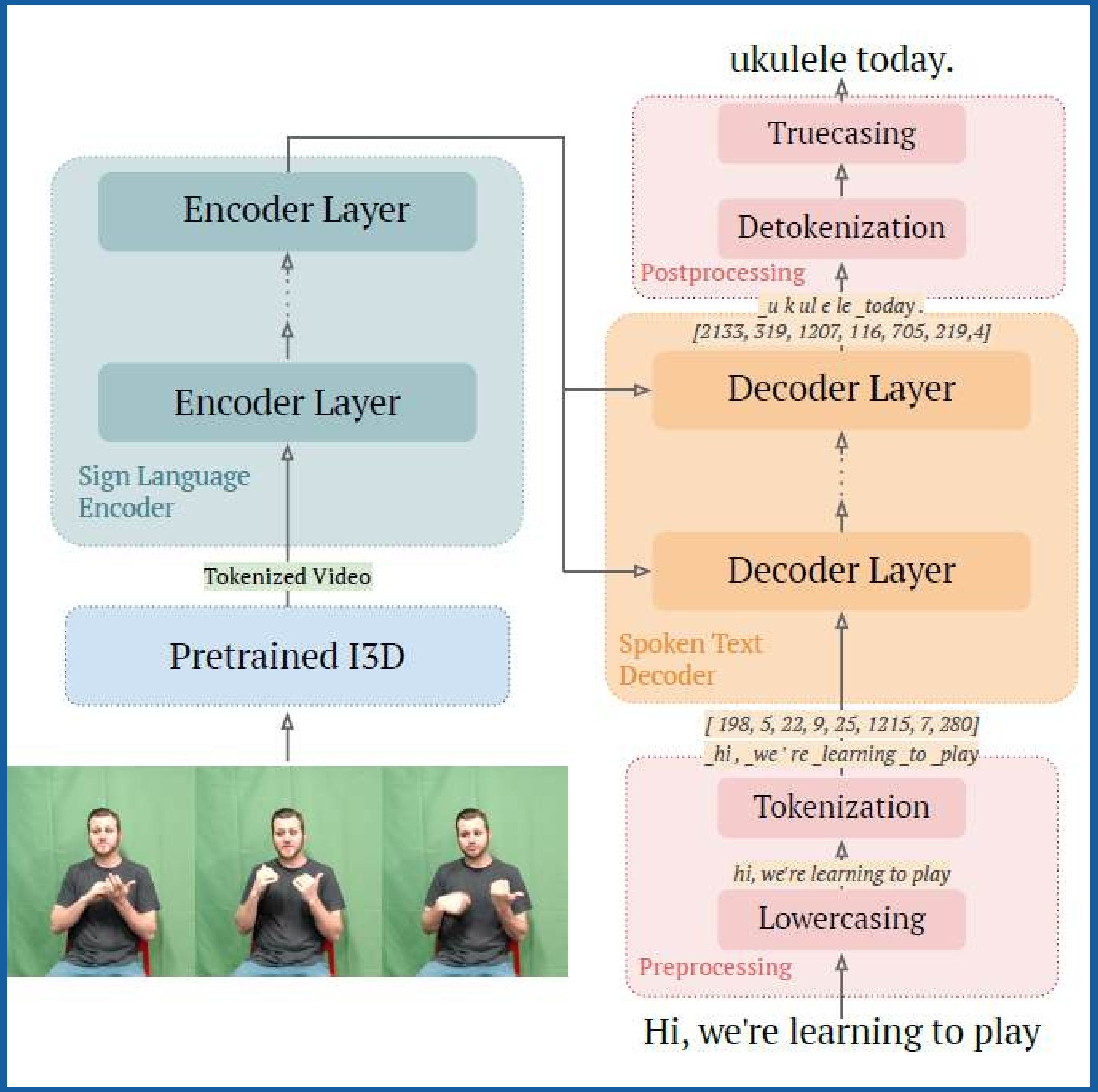
MS-ASL

I3D (Standford)

Sign2Gloss2Text

# Approach 2

How2Sign




New Model

Tarrés, L., Gállego, G. I., Duarte, A., Torres, J., & Giró-i-Nieto, X. (2023). Sign language translation from instructional videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5625-5635).

# Approach 2

- Video data passed through pretrained I3D to extract spatial-temporal features, saved as numpy files.
- Transformer layers process the features, generating English text translations.

### Advantages

- Model focuses on critical video segments, improving contextual accuracy in sentence generation.

### Disadvantages

- Low BLEU score, limiting practical use.
- Model struggles with complex, lengthy sign sequences.
- The lengthy training time hinders model improvement and testing efficiency.

# Approach 3

- Sign2Signwriting2Text
- Text2Signwriting2Sign

We have taken a different approach as a result of these constraints, and this is the strategy we decided to use for our project: employing SignWriting as a translation intermediate. SignWriting, as instead of gloss, is a writing system designed especially for sign languages. More accurate translation is made possible by our method's use of pretrained models. This method's ability to achieve a higher BLEU score than the prior model is one of its main advantages. Section 4 provides more information on this strategy.

# 4. Experiment

## 4.1 System

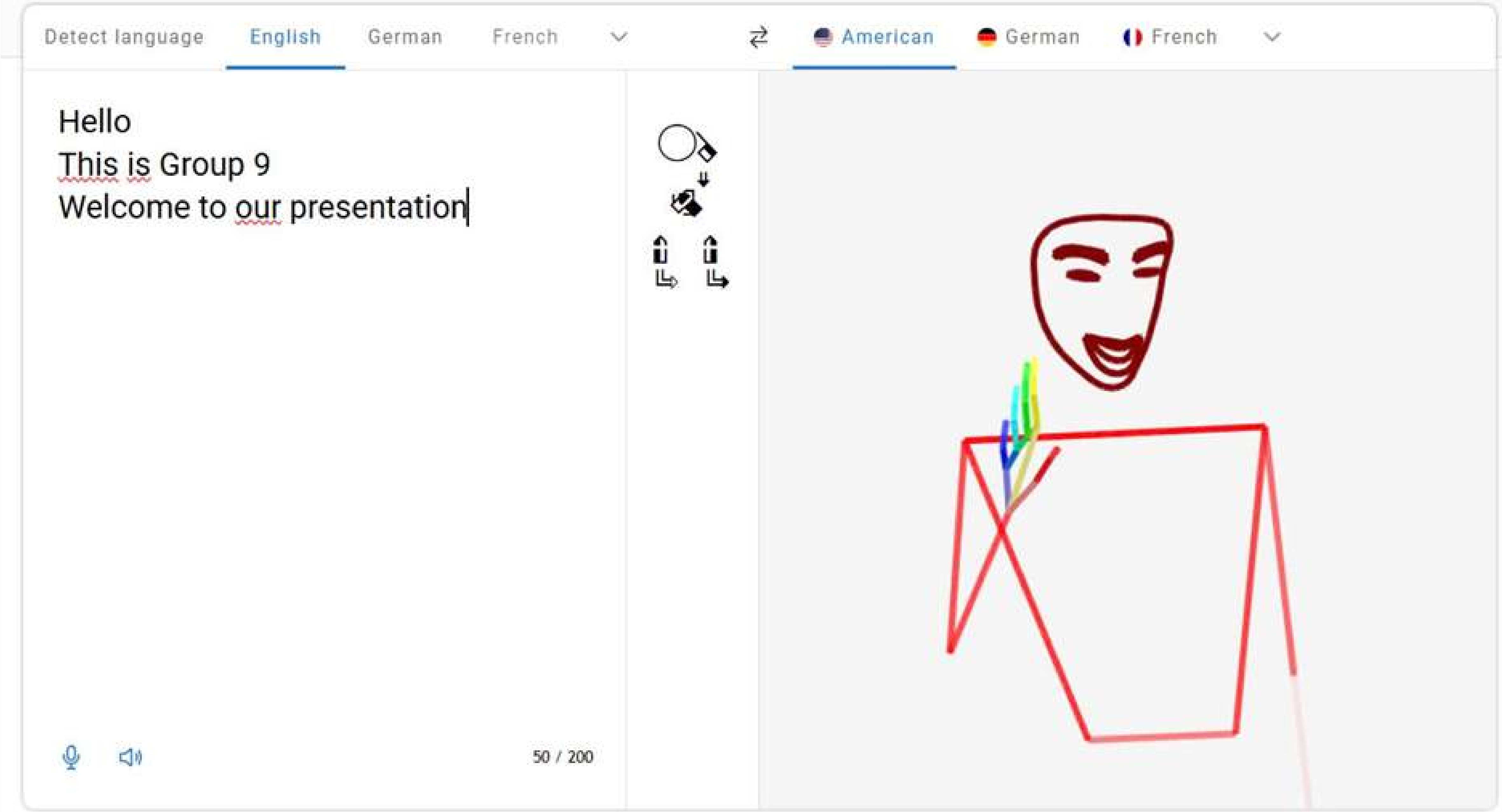
TS

大 / T

{ TRANSLOCO }

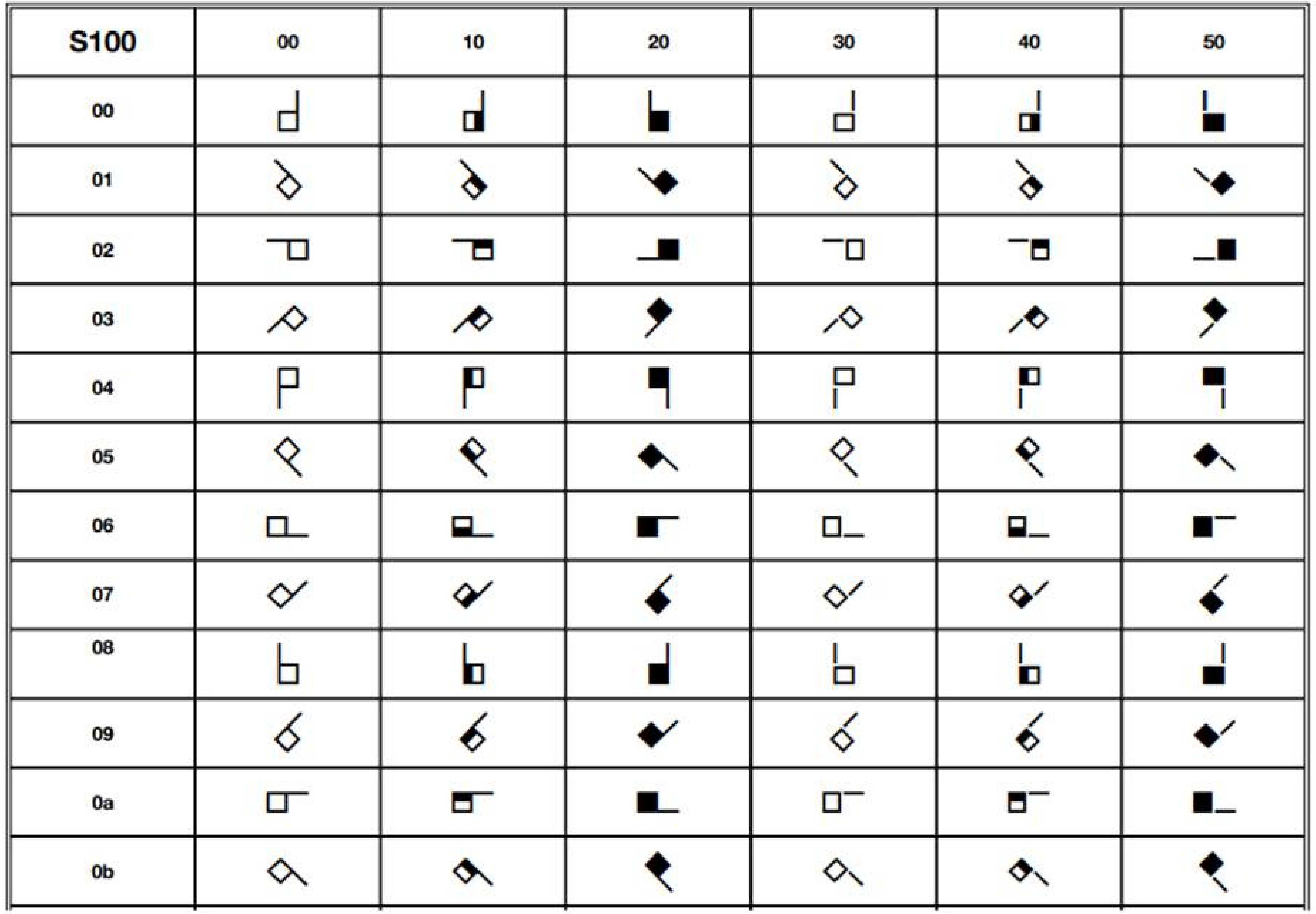
MediaPipe

# 4.2 AI Architecture

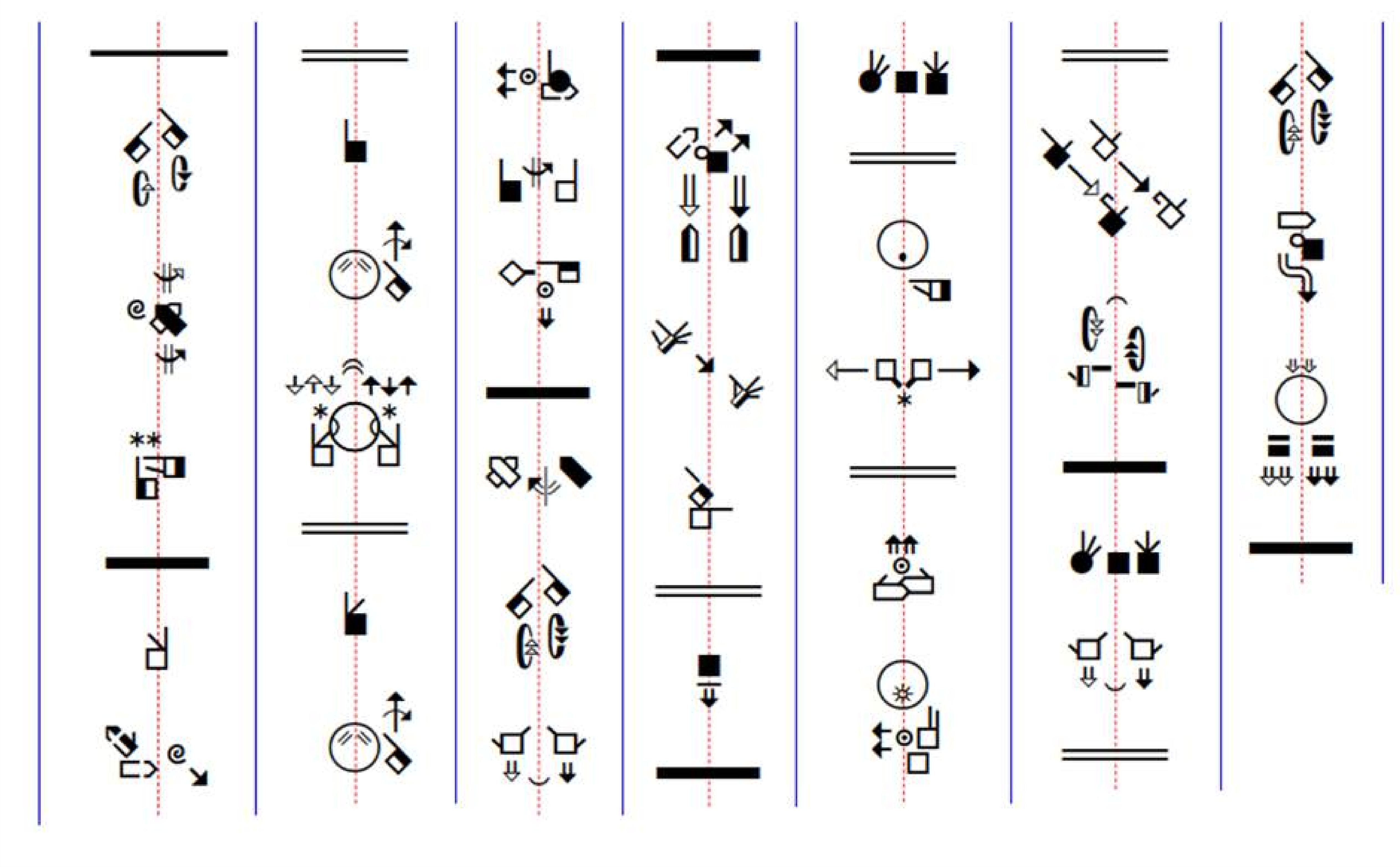



**Demo application based on our models, translating from spoken languages to signed languages represented in SignWriting, then to human poses.**

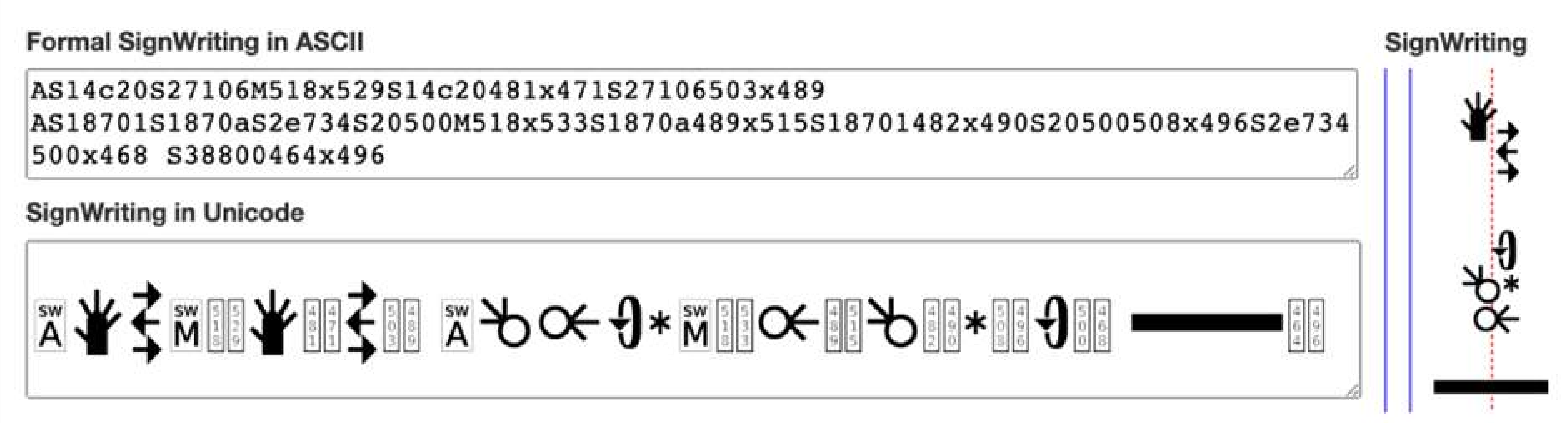
Hand shapes and their equivalents in SignWriting




Orientation of a symbol in SignWriting in 3D space. Each row applies a rotation of the palm in a 2D space vertical to the ground. Each column applies a rotation of the palm in a 2D space parallel to the ground. This can be seen as a factorization of the symbol S100xx to its core S100 plus row and column numbers

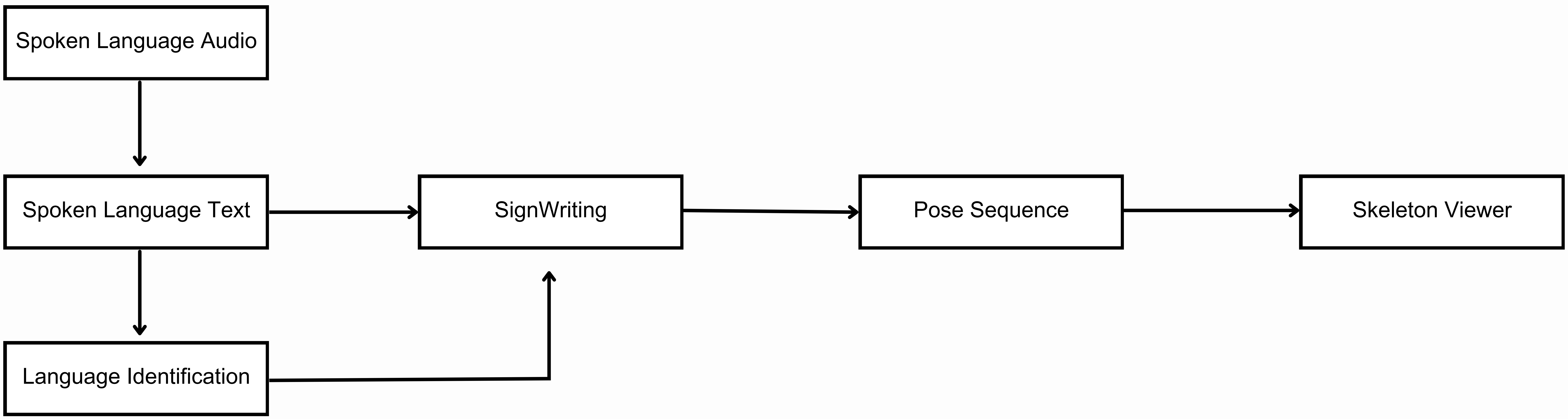


An example of SignWriting written in columns, ASL translation of an introduction to Formal SignWriting in ASCII. The relative positions of the symbols within the box iconically represent the locations of the hands and other parts of the body involved in the sign being represented



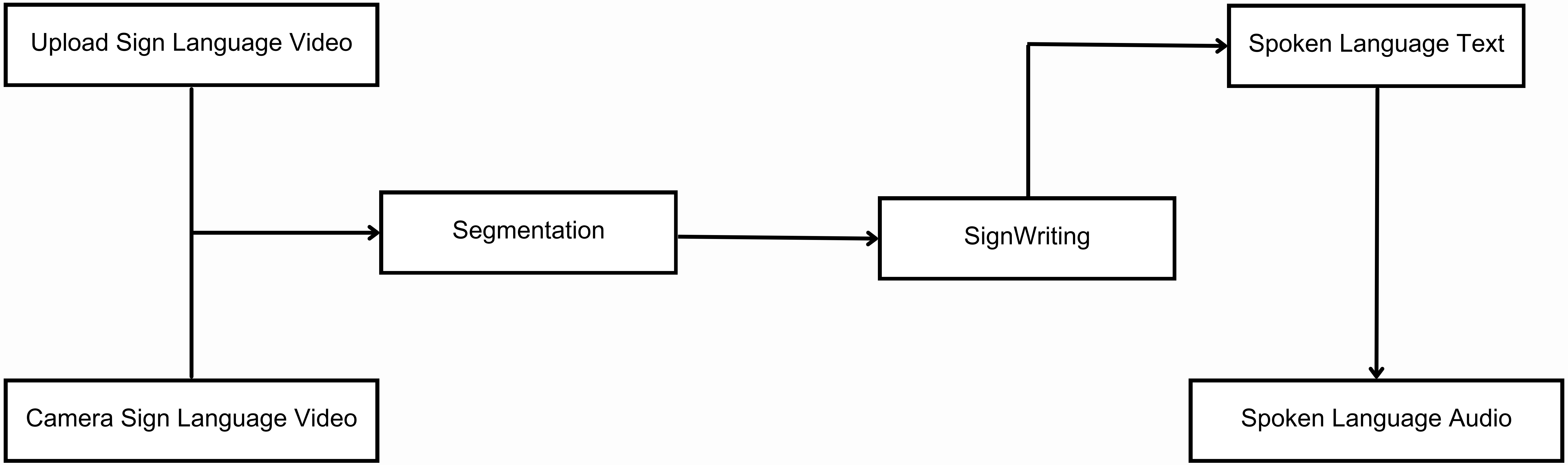

“Hello world.” in FSW, SWU and SignWriting graphics. In FSW/SWU, A/SWA and M/SWM are the box markers (acting as sign boundaries); S14c20 and S27106 (graphemes in SWU) are the symbols; 518 and 529 are the x, y positional numbers on a 2-dimensional plane that denote symbols’ position within a sign box, S38800 (horizontal bold line in SWU) is the punctuation full stop symbol.

# Details on how to build the model




**Spoken language text into Skeleton Viewer**

# Details on how to build the model

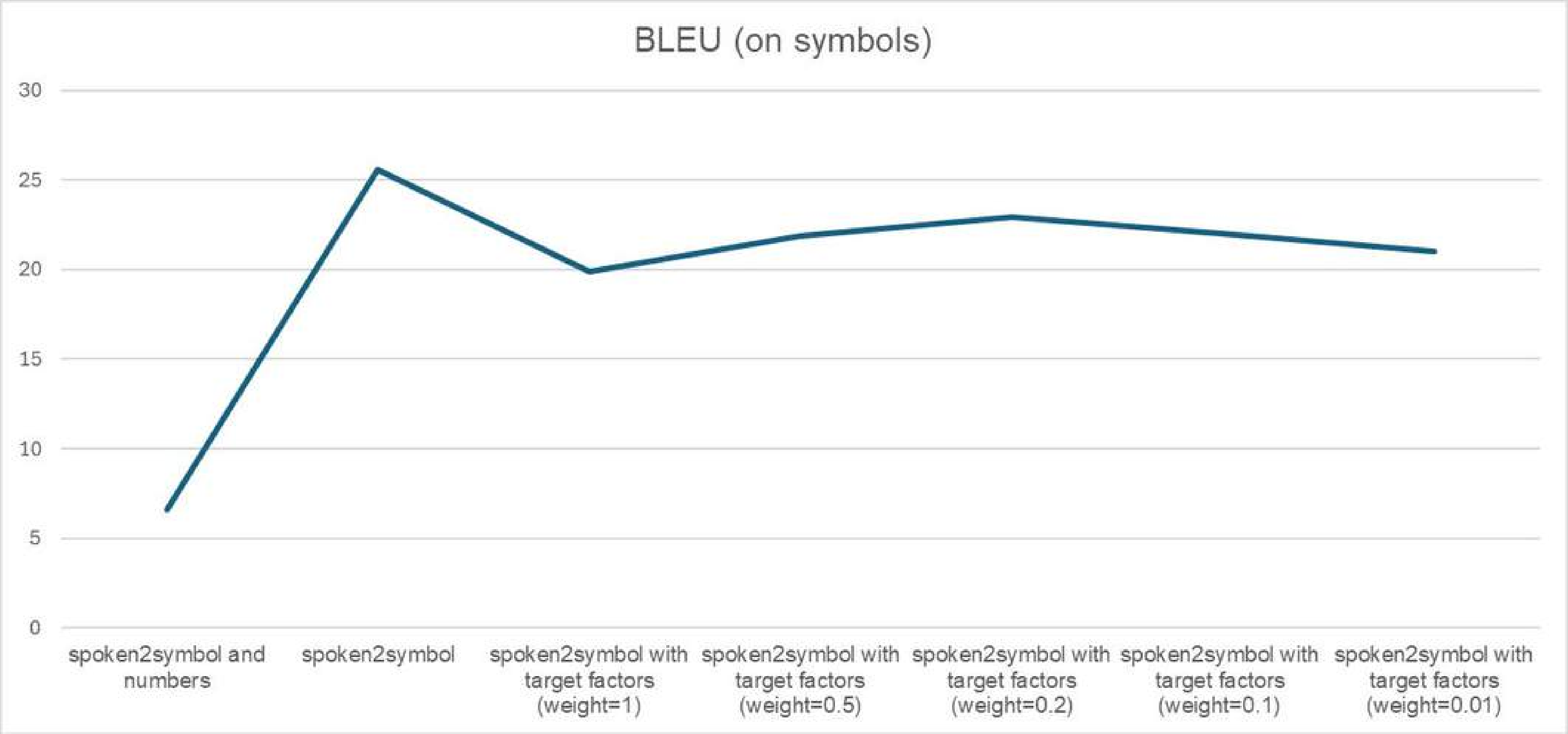


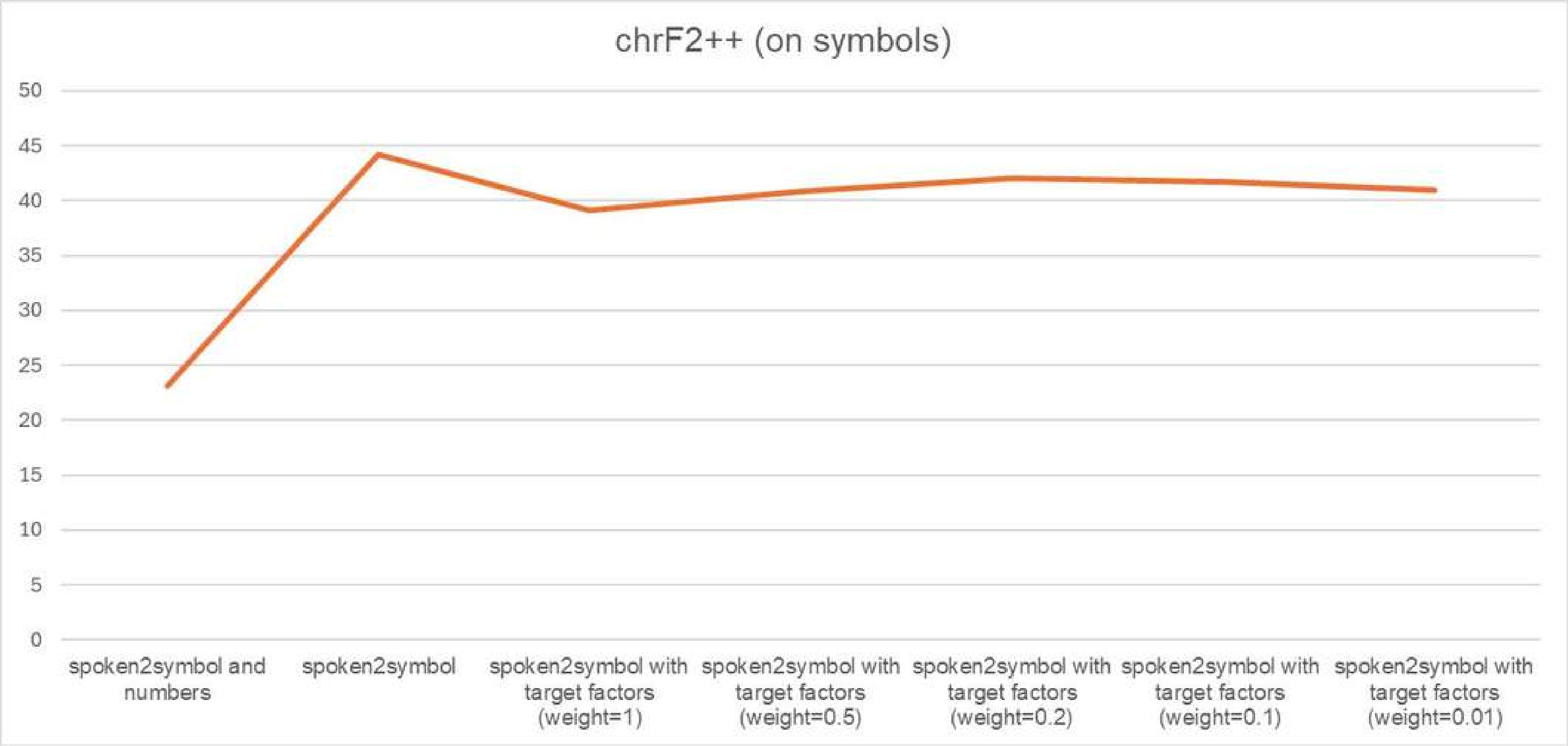

**Sign language videos into spoken language text and audio output**

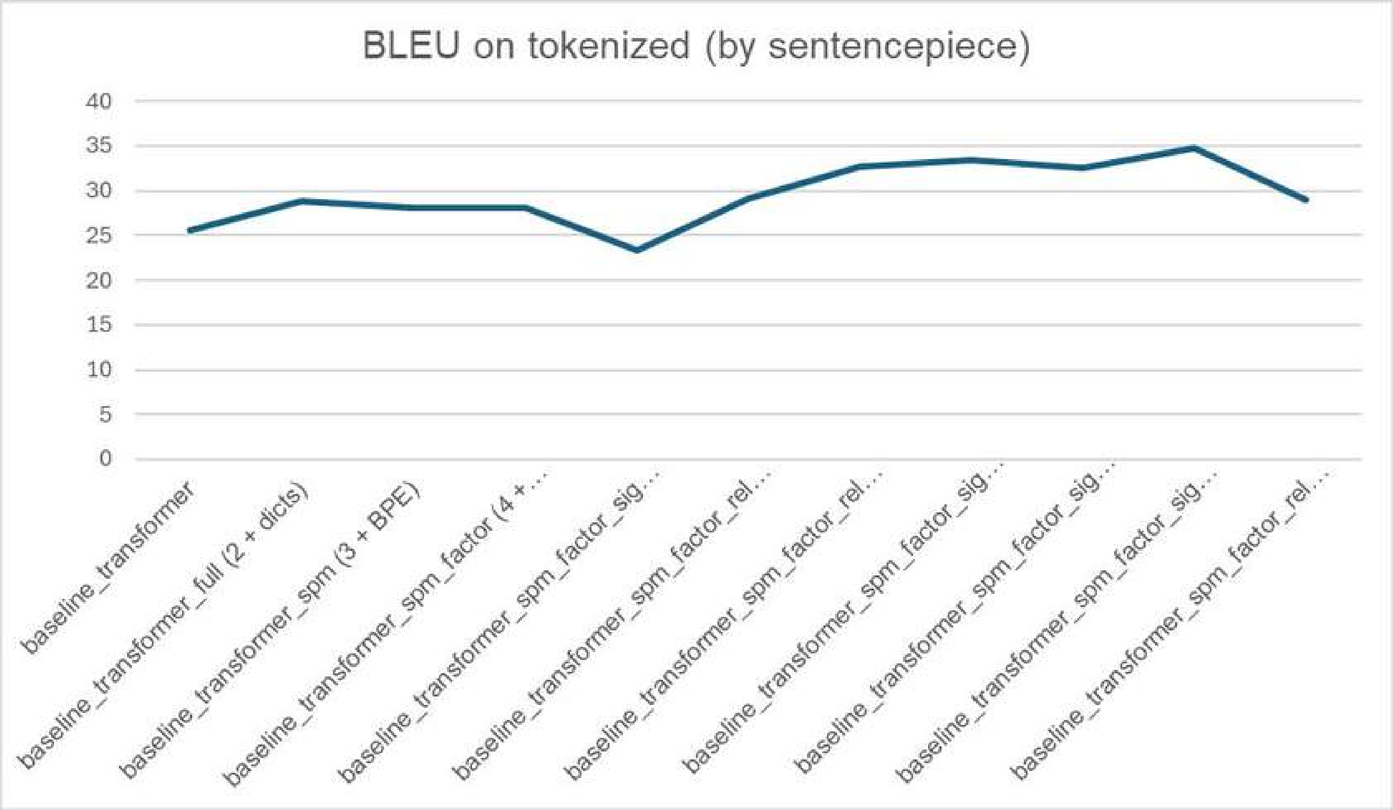
# 5. Results and Analysis

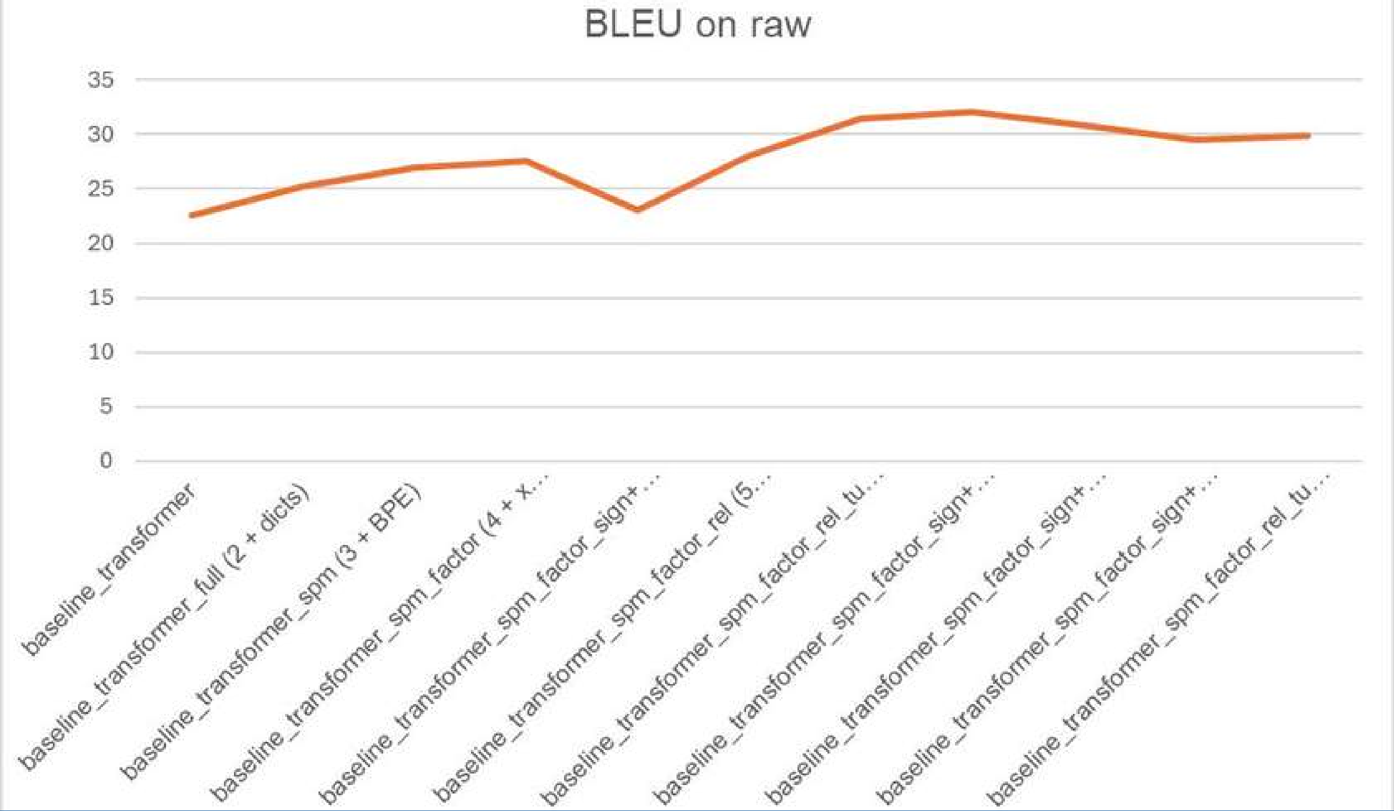
5.1 Results

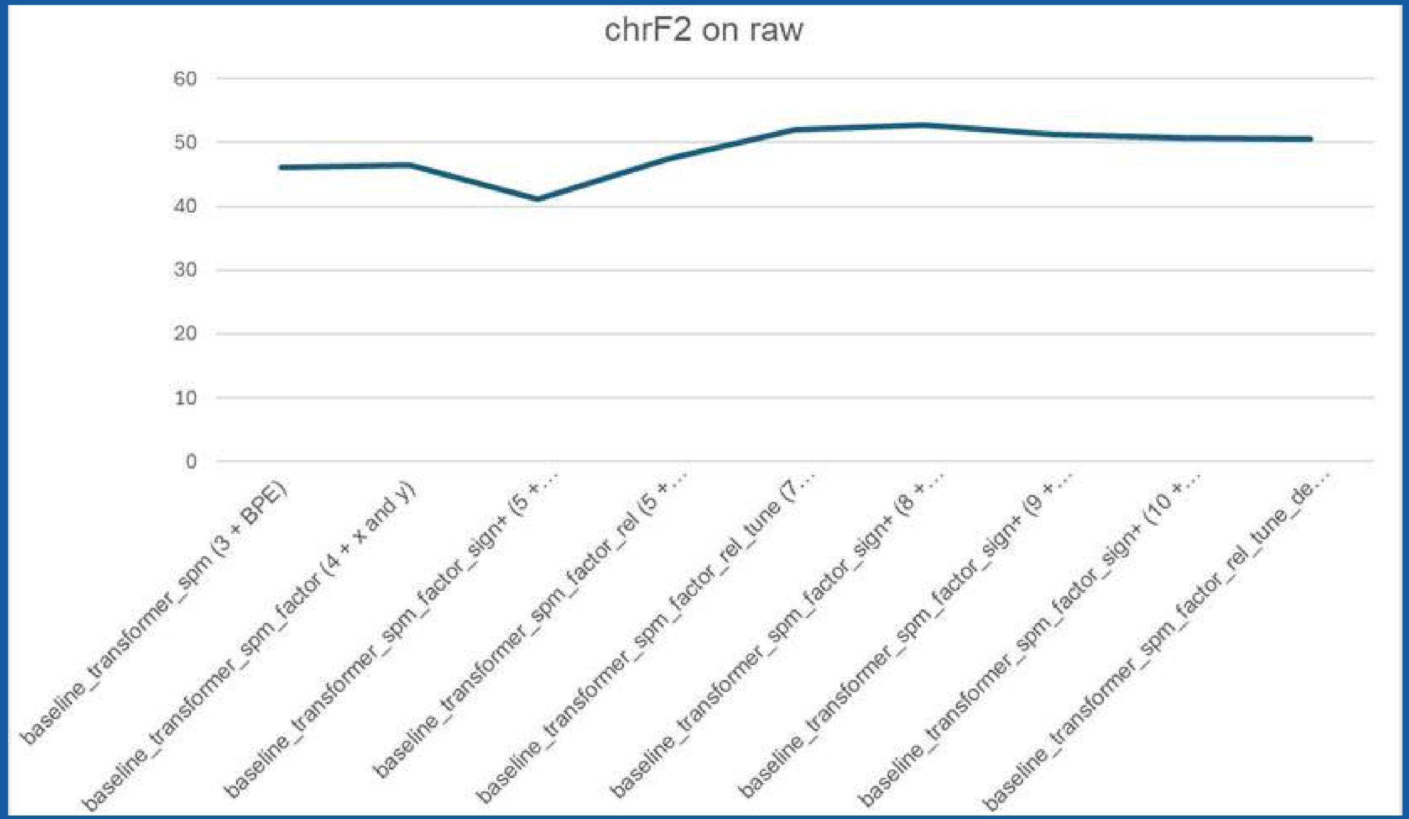
5.2 Analysis


BLEU (on symbols)
30
25
20
15
10
5
0
spoken2symbol and numbers
spoken2symbol
spoken2symbol with target factors (weight=1)
spoken2symbol with target factors (weight=0.5)
spoken2symbol with target factors (weight=0.2)
spoken2symbol with target factors (weight=0.1)
spoken2symbol with target factors (weight=0.01)


chrF2++ (on symbols)
50
45
40
35
30
25
20
15
10
5
0
spoken2symbol and numbers
spoken2symbol
spoken2symbol with target factors (weight=1)
spoken2symbol with target factors (weight=0.5)
spoken2symbol with target factors (weight=0.2)
spoken2symbol with target factors (weight=0.1)
spoken2symbol with target factors (weight=0.01)


BLEU on tokenized (by sentencepiece)
40
35
30
25
20
15
10
5
0
baseline_transformer
baseline_transformer_full (2 + dicts)
baseline_transformer_spm (3 + BPE)
baseline_transformer_spm_factor (4 +...
baseline_transformer_spm_factor_sig...
baseline_transformer_spm_factor_rel...
baseline_transformer_spm_factor_rel...
baseline_transformer_spm_factor_sig...
baseline_transformer_spm_factor_sig...
baseline_transformer_spm_factor_sig...
baseline_transformer_spm_factor_rel...


BLEU on raw
35
30
25
20
15
10
5
0
baseline_transformer
baseline_transformer_full (2 + dicts)
baseline_transformer_spm (3 + BPE)
baseline_transformer_spm_factor (4 + x...
baseline_transformer_spm_factor_sign+...
baseline_transformer_spm_factor_rel (5...
baseline_transformer_spm_factor_rel_tu...
baseline_transformer_spm_factor_sign+...
baseline_transformer_spm_factor_sign+...
baseline_transformer_spm_factor_sign+...
baseline_transformer_spm_factor_rel_tu...


chrF2 on raw
60
50
40
30
20
10
0
baseline_transformer_spm (3 + BPE)
baseline_transformer_spm_factor (4 + x and y)
baseline_transformer_spm_factor_sign+ (5 +...
baseline_transformer_spm_factor_rel (5 +...
baseline_transformer_spm_factor_rel_tune (7...
baseline_transformer_spm_factor_sign+ (8 +...
baseline_transformer_spm_factor_sign+ (9 +...
baseline_transformer_spm_factor_sign+ (10 +...
baseline_transformer_spm_factor_rel_tune_de...

| No. | Model | Type | Data | BLEU-4 | CHRF |
| --- | --- | --- | --- | --- | --- |
| Approach 1 | Sign2Gloss2Text | Translate Video Sign Language to Text | MS-ASL | - | - |
| Approach 2 | basic Transformer seq2seq endocder-decoder (baseline) | Translate Video Sign Language to Text | How2Sign | 8.43 | 9.6 |
| Approach 3 | spoken2symbol | Translate Video Text to Sign Language | SignBank+ | **25.6** | **44.2** |
| | baseline_transformer_spm_factor_sign+ (10 + smaller bpe vocab 2000 -> 1000) | Translate Video Sign Language to Text | SignBank+ | **34.8** | **50.8** |

# 6. Conclusion and Future Work

6.1 Conclusion

6.2 Future Work

# 6.1. Conclusion

The current project has achieved significant results in developing a sign language translation system, enabling the conversion of text to sign language and vice versa from sign language to text for three countries: the United Kingdom, France, and Germany. Furthermore, the project has expanded to include additional sign languages, although this expansion is still incomplete.

### Advantages

- The feasibility of the project in developing a communication support tool for the deaf community.
- Provides translation for sign languages from different countries.

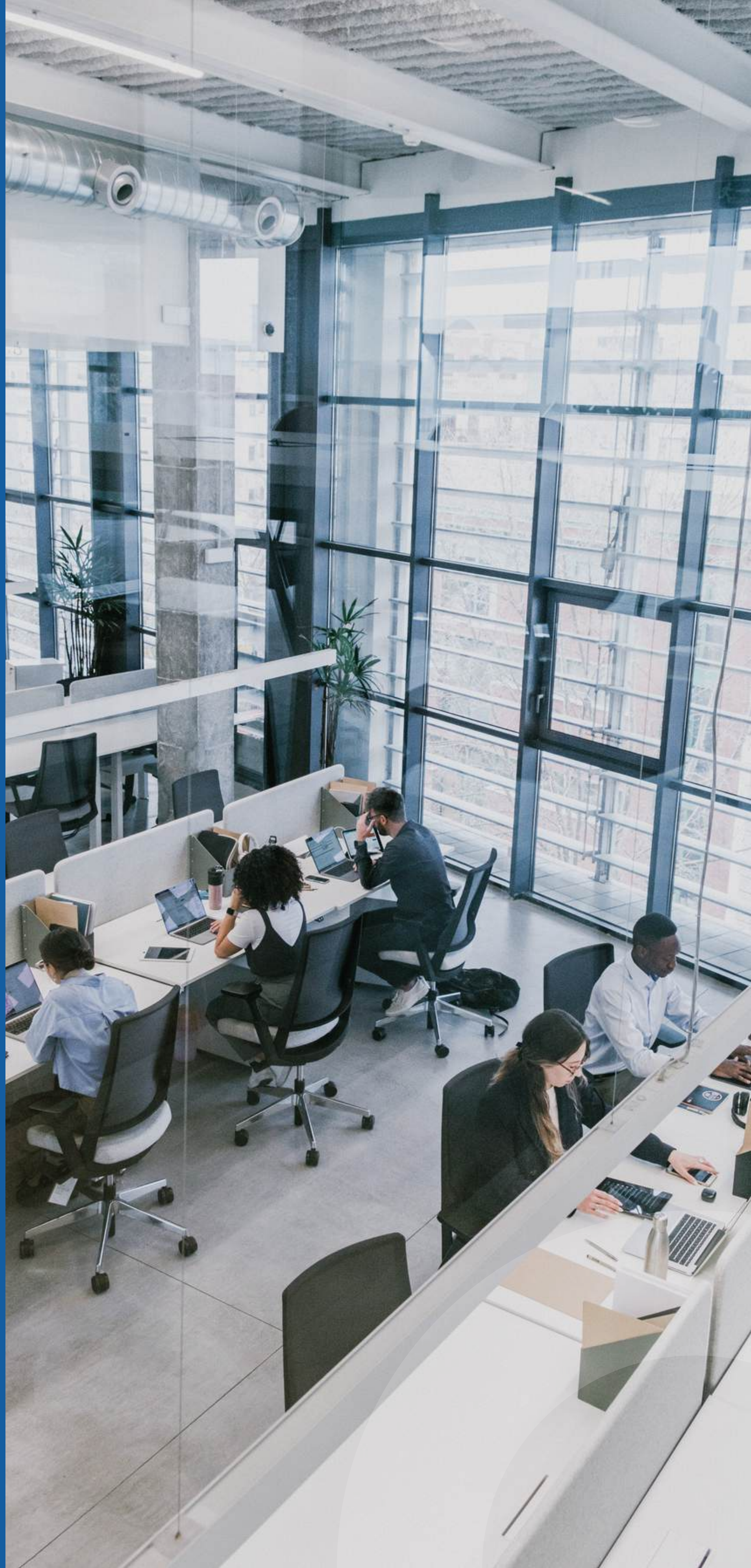
### Disadvantages

- The grammatical structure of the sign language is not yet entirely accurate.

# 6.2. Future Work

Enhance translation services to support multiple languages.

Improve grammatical accuracy in sign language translation.



# THANK YOU FOR YOUR ATTENTION!