

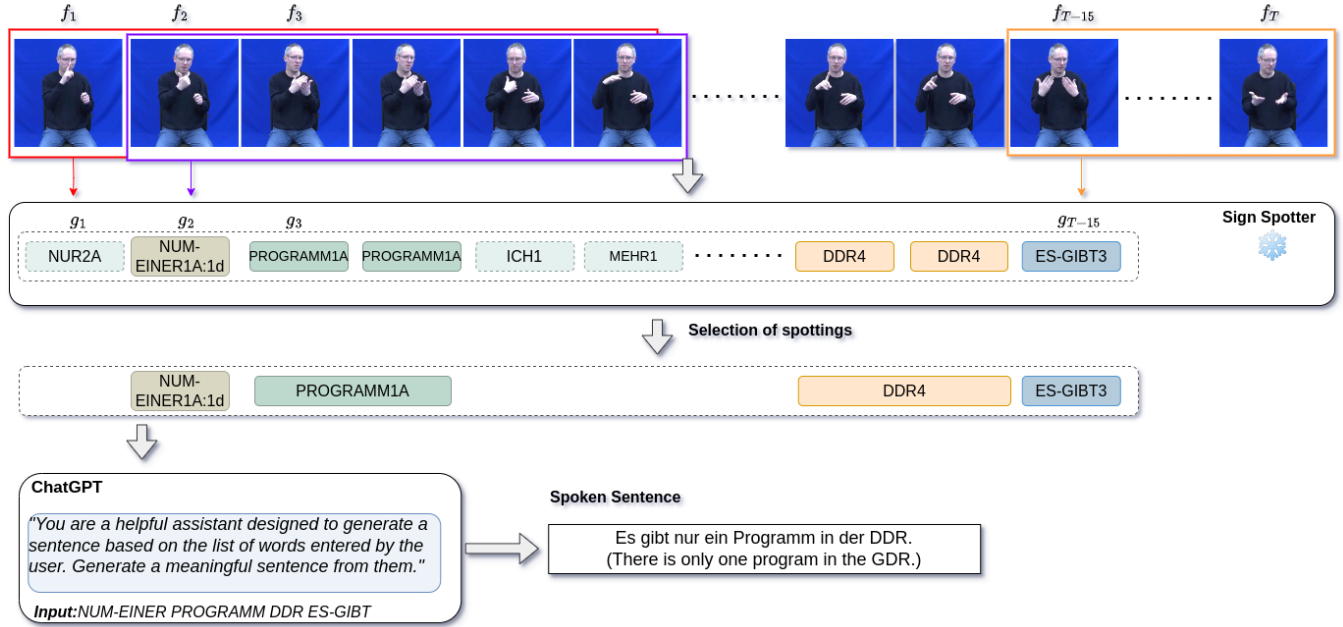
# Spotter+GPT: Turning Sign Spottings into Sentences with LLMs

Ozge Mercanoglu Sincan  
CVSSP

University of Surrey  
Guildford, Surrey, United Kingdom  
o.mercanoglusincan@surrey.ac.uk

Richard Bowden  
CVSSP

University of Surrey  
Guildford, Surrey, United Kingdom  
r.bowden@surrey.ac.uk



**Figure 1: Overview of the proposed sign language translation framework. The system spots signs by processing video input and generates spoken language sentences via ChatGPT.**

## Abstract

Sign Language Translation (SLT) is a challenging task that aims to generate spoken language sentences from sign language videos. In this paper, we introduce a lightweight, modular SLT framework, Spotter+GPT, that leverages the power of Large Language Models (LLMs) and avoids heavy end-to-end training. Spotter+GPT breaks down the SLT task into two distinct stages. First, a sign spotter identifies individual signs within the input video. The spotted signs are then passed to an LLM, which transforms them into meaningful spoken language sentences. Spotter+GPT eliminates the requirement for SLT-specific training. This significantly reduces

computational costs and time requirements. The source code and pretrained weights of the Spotter are available online <sup>1</sup>.

## CCS Concepts

• **Human-centered computing** → **Accessibility technologies.**

## Keywords

Sign Spotting, Sign Language Translation, Real-time, ChatGPT

### ACM Reference Format:

Ozge Mercanoglu Sincan and Richard Bowden. 2025. Spotter+GPT: Turning Sign Spottings into Sentences with LLMs. In *ACM International Conference on Intelligent Virtual Agents (IVA Adjunct '25)*, September 16–19, 2025, Berlin, Germany. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3742886.3756708>

## 1 Introduction

Sign languages are visual languages that rely on manual hand articulations, facial expressions, and body movements. To bridge communication gaps between the Deaf community and hearing people, Sign Language Translation (SLT) (sign language → spoken

<sup>1</sup><https://gitlab.surrey.ac.uk/cogvispublic/sign-spotter>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

IVA Adjunct '25, Berlin, Germany

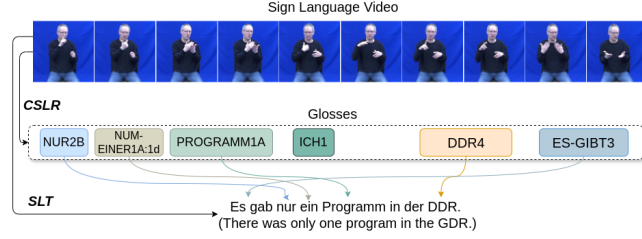
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-1996-7/25/09

<https://doi.org/10.1145/3742886.3756708>

language) and Sign Language Production (SLP) (spoken language  $\rightarrow$  sign language) systems hold great significance.

SLT is often formulated as a Neural Machine Translation (NMT) task since it aims to generate spoken/written language sentences from sign language videos [5]. Compared to classical text-based NMT approaches, which work on easily tokenizable text, SLT deals with continuous sign language videos, which are hard to align and tokenize. Furthermore, spoken and sign languages have different grammar, and the order of the sign glosses and spoken language words are different as seen in Fig. 2.



**Figure 2: Overview of the CSLR (Continuous Sign Language Recognition) and SLT (Sign Language Translation) tasks.**

To address this issue, some researchers have approached SLT as a combination of two sub-tasks; Sign Language Recognition (SLR), recognizing the constituent signs of sequences and then translating the recognized signs into meaningful spoken language sentences [5, 6, 42]. Some researchers approach the first task as Continuous Sign Language Recognition (CSLR) and try to detect sequences of glosses as an intermediate representation to represent sign language videos. Camgoz et. al [5, 6] showed that using gloss-based intermediate representations improved SLT performance significantly. The common approach in CSLR is learning spatial and temporal visual representations with a sequence-to-sequence Connectionist Temporal Classification (CTC) loss [12]. These CSLR approaches require gloss supervision, which is hard and time-consuming to accurately annotate.

On the other hand, some researchers tackle SLT in an end-to-end manner by training SLT models that produce spoken language sentences directly from sign language videos [6, 41, 44]. Although end-to-end approaches can achieve excellent results on small SLT datasets, such as PHOENIX-2014-T [5], they under perform on SLT datasets that are weakly aligned or have a large domain of discourse [2, 20, 32].

In this work, we explore an alternative, modular approach: Spotter+GPT. Our goal is to reduce the cost and complexity of SLT pipelines by leveraging pre-trained Large Language Models (LLMs) instead of training a gloss-to-text translation model. We first train a sign spotter using a large sign language dataset from the linguistic domain. Then, we utilize our spotter to recognize the sequence of sign glosses from continuous sign language videos. These glosses are then passed to the LLM via prompting to generate spoken sentences. We evaluate our method on the MeineDGS dataset and a custom DGS-20 benchmark and compare it to traditional gloss-to-text transformer baselines. This pipeline does not require training an SLT-specific decoder and can generate semantically coherent translations, particularly in controlled settings.

## 2 Related Work

### 2.1 Sign Language Recognition

SLR can be divided into two categories; Isolated Sign Language Recognition (ISLR) [1–3, 15, 17, 33, 39] and Continuous Sign Language Recognition (CSLR) [9, 25, 38, 43]. ISLR focuses on recognizing a single sign from the video, while CSLR focuses on recognizing the sequence of glosses. CSLR is a weakly supervised recognition task since continuous SLR datasets [5, 10] usually provide gloss sequences without explicit temporal boundary since labeling each gloss frame by frame is a time-consuming process. For this reason, using the CTC loss [12] became popular [6, 25, 38, 43].

Some researchers develop CSLR models with the assistance of an ISLR model [9, 38]. Cui et. al [9] first trained a feature extraction module and then finetuned the whole system iteratively. Wei et. al [38] trained two ISLR models in two different sign languages and then proposed a multilingual CSLR by utilizing cross-lingual signs with their assistance. In this work, we utilized an ISLR model to detect the sequence of glosses.

### 2.2 LLMs for Sign Language Translation

The development of Large Language Models (LLMs) has led to significant improvements in natural language processing tasks [4, 27, 35]. After OpenAI introduced ChatGPT, which was trained using Reinforcement Learning from Human Feedback (RLHF), it has attracted significant attention because of its success in several tasks, including question answering, summarization, machine translation, etc [19, 22, 26].

Despite the effectiveness of ChatGPT, only a few studies have explored ChatGPT in the context of sign language. Shahin and Ismail [31] evaluated ChatGPT’s capabilities on gloss-to-text and text-to-gloss translations with a limited set of phrases, such as with only 5 medical-related statements in English and Arabic sign-spoken language pairs. While promising, their experiments were minimal and relied on given gloss inputs rather than real video inputs. They found that it performs better when translating to English rather than Arabic.

More recently, several studies have begun integrating LLMs into SLT pipelines, often through training adapters or fine-tuning [8, 11, 13, 18, 40]. However, such systems typically require additional training and resources, limiting their ease of deployment.

In contrast, our approach, Spotter+GPT, is a lightweight SLT framework that uses a pretrained sign spotter to detect glosses from video, and directly prompts ChatGPT for spoken sentence generation. To the best of our knowledge, it is the first to combine gloss spotting from raw video with prompt-based LLM translation, without requiring any additional training or adaptation for translation.

## 3 Method

As shown in Fig. 1, our approach contains a two-step process. First, we train a sign spotter to identify individual signs. This model was trained to recognize a single isolated sign from the input video. Once trained, the spotter is applied to continuous sign language videos in a sliding window manner to get gloss predictions. Then we pool and threshold these detections to obtain the final set of

identified glosses. We call this first step Sign Spotting (Section 3.1). Subsequently, we employ a pretrained LLM to convert these sequences of glosses into spoken language sentences. To achieve this, we prompted a ChatGPT model (Section 3.2).

### 3.1 Sign Spotting

As spotter, we employ an I3D model [7]. We prepare isolated sign sequences from a continuous sign language dataset, MeineDGS [16], where the glosses are annotated in frame level. Within the MeineDGS dataset the average duration of a gloss is 10 frames while we train our I3D models with a window size of 16. If an isolated sign sequence is shorter than 16 frames, we repeat the last frame. For longer sign instances we generate multiple samples by using a sliding window with a stride of 8.

In the training time, the model takes 16 consecutive frames. We resize the input images to  $256 \times 256$  and then crop to a  $224 \times 224$  region. We replaced the ReLU activation function with the Swish activation function [28] as it improves sign language recognition performances [14, 39].

Following [2], we utilize cross-entropy loss and an SGD optimizer [34] with momentum 0.9, batch size 4, and an initial learning rate of 0.01. We decreased the learning rate by a factor of 10 when validation loss plateaus for 4 consecutive epochs. We use label smoothing of 0.1 to prevent overfitting. While our input videos are cropped randomly during training time, it is cropped from the center during evaluation. We applied color augmentation during training.

After training our I3D model, we employ it to spot signs in coarticulated continuous sign language videos in a sliding window manner. With a stride of 1, for the given input video with  $T$  frames, we obtain  $T - 15$  gloss predictions. Then, we propose a straightforward yet efficient solution to produce a final sequence of glosses. First, we filter gloss predictions by a threshold, based on the model’s prediction confidence. Following this, we collapse the consecutive repeated predictions to obtain the final sequence of gloss predictions.

### 3.2 ChatGPT

Due to the success of ChatGPT in many tasks, we utilize it to create spoken sentences from glosses. This strategy not only eliminates the requirement for gloss-to-text model training but also offers potential advantages. ChatGPT can generate fluent and context-aware sentences, making it a strong candidate for translating glosses into coherent sign language sentences.

It is worth noting that the order of the glosses in sign languages are different from the word order in spoken languages. Meanwhile, the vocabulary of our gloss spotter is constrained by the number of classes on which the I3D model was trained on. However, the dynamic nature of ChatGPT allows a flexible translation process, enabling us to overcome the limitations of a fixed gloss set and, different ordering.

We prompt ‘gpt-3.5-turbo’ version of ChatGPT via the OpenAI Python API without fine-tuning it. For each input video, after we obtain a series of glosses, we pass them to ChatGPT and prompt the model to create a meaningful sentence using those glosses.

The spotter is capable of real-time inference on systems equipped with a GPU. We tested it on two machines: a desktop with an NVIDIA RTX 3090 and a laptop with an NVIDIA RTX 5000. In both cases, the system operated smoothly, achieving a minimum of 25 frames per second. Since ChatGPT is accessed via an external API, its response time depends on network conditions. In our observations, the typical latency per prompt was approximately 1-2 seconds.

### 3.3 Prompt Engineering

We initially set our prompt by only defining our task as generating a sentence based on the list of words entered by the user. However, we observed that sometimes ChatGPT produces outputs like “Sorry, I could not generate a sentence.” in cases when spotter failed to detect any gloss or when the number of detected glosses was insufficient. To address this, we refined the prompt by adding two explicit rules to avoid unrelated sentences and to produce “No translation.” in cases where a meaningful sentence could not be generated. The final prompt is as follows:

- You are a helpful assistant designed to generate a sentence based on the list of words entered by the user. You need to strictly follow these rules:
  - (1) The user will only give the list of German words separated by a space, you just need to generate a meaningful sentence from them.
  - (2) Only provide a response containing the generated sentence. If you cannot create a German sentence then respond with “No Translation”.

## 4 Experiments

### 4.1 Dataset and Preprocessing

**MeineDGS.** It is a large linguistic German Sign Language (DGS) dataset [16]. Videos contain free-flowing conversation between two deaf participants. We follow the sign language translation protocol set [29] on MeineDGS, which has 40,230 training, 4,996 development, and 4,997 test sentences.

We use the MeineDGS-V split [29], which distinguishes between sign variants, each containing the same meaning but with differing motions. MeineDGS-V has approximately 10,000 glosses available for training. However, some signs have relatively few examples or some of them are singletons. To make the dataset more balanced for isolated SLR model training, we selected glosses that have more than 12 samples and we exclude the *INDEX* gloss as it stands for a pointing, making it the predominant gloss in the dataset. This criterion leads to 2,301 classes. We train our I3D spotter on these 2,301 classes using train, validation, and test splits of Saunders et al [29].

Please note that in [29] the spoken sentences are lowercase, punctuation was removed, and all German characters (ä, ö, ü, and ß) were replaced with corresponding English letters. This will cause some words to change meaning. Therefore, instead of the sentence representation of [29], we used the original spoken language sentences [16] in the sign language translation task.

**DGS-20 Videos.** To further evaluate our approach, we collected a small dataset containing 20 German Sign Language (DGS) videos. These videos were recorded by a single deaf signer, who performed

10 unique sentences, each repeated twice for consistency and variation. The glosses used in these sentences were selected from the 2,301-class vocabulary of our trained sign spotter, ensuring full coverage. Some example sentences are presented in Table 1.

**Table 1: Some examples from DGS-20 dataset.**

ID	German Sentence (English Translation)
1	Die Familie isst abends im Restaurant. (The family eats in the restaurant in the evening.)
2	Der Dolmetscher spricht mit der Familie des Mädchens. (The interpreter speaks to the girl's family.)
3	Der Dozent fährt morgens mit dem Fahrrad zur Universität. (Tomorrow the lecturer will ride his bike to the university.)
4	Der Schauspieler spricht viel über Politik und Kultur. (The actor talks a lot about politics and culture.)

## 4.2 Evaluation Metrics

We use BLEU [21], and BLEURT [30] metrics to evaluate the performance of our SLT approach. BLEU is a metric based on the precision of n-grams (consecutive word sequences) for machine translation. On the other hand, BLEURT aims to achieve human-quality scoring. Higher scores indicate better translation. We use the sacreBLEU [23] implementation for BLEU, and BLEURT-20 checkpoints [24] for the calculation of BLEURT scores.

## 4.3 Quantitative Results

**Sign Language Recognition:** We evaluate our I3D model on the MeineDGS dataset providing the per-instance and per-class accuracy scores. We first fine-tuned I3D pretrained on Kinetics [7] and obtained 53.24% and 40.70%, respectively. It has been shown that pretraining on larger sign recognition datasets improves sign performance [1, 37]. Therefore, we first fine-tuned the I3D model on the large-scale BOBSL (BBC-Oxford British Sign Language) dataset [2]. Then we fine-tuned an I3D model pretrained on Kinetics + BOBSL for finetuning on the MeineDGS. We observe 1.5% performance increase with the usage of sign language data in weight initialization. Our results are provided in Table 2.

**Table 2: Isolated sign language recognition performance on MeineDGS-V.**

Model	Pretrained on	Per-instance	Per-class
I3D	Kinetics [7]	53.24%	40.70%
I3D	Kinetics [7] + BOBSL [2]	<b>54.57%</b>	<b>42.48%</b>

**Sign Language Translation:** We evaluate our entire SLT approach, Spotter+GPT, on the MeineDGS-V test split and our DGS-20 videos. The quantitative results are provided in Table 3 and Table 4, respectively.

We empirically evaluate the Spotter's performance using varying probability thresholds. The threshold of 0.7 for the probability associated with each gloss prediction yielded the best results.

It is worth mentioning that DGS-20 videos result in higher performance than MeineDGS. As this dataset contains a limited and controlled vocabulary, GPT achieves significantly higher scores in both BLEU and BLEURT metrics, demonstrating its effectiveness in low-resource but constrained setups.

**Evaluation of each component.** To evaluate the performance of our components independently we conduct two different types of experiments. First, to evaluate the performance of our spotter, we replace its results with the ground truth gloss annotations. Although the MeineDGS-V test split has 4,620 glosses, our spotter is only able to recognize 2,301 glosses. To make a fair comparison, we excluded gloss annotations that do not belong to our Spotter's vocabulary and we refer to this filtered reference set as Sub-GT. Specifically, Sub-GT is derived from the full ground truth annotations in the MeineDGS-V, but only includes glosses that exist in our predefined vocabulary. We refer to our full pipeline using this subset as Sub-GT+GPT. As expected, these scores surpass the results obtained with the spotter (BLEURT: 29.72 vs. 21.62).

Second, to evaluate the role of ChatGPT as a gloss-to-text generator, we replace it with a traditional Transformer model [36], trained on gloss-sentence pairs. We use two layers with 8-heads in the transformer encoder and decoder using 512 hidden units. We use the Adam optimizer with an initial learning rate of  $6 \times 10^{-4}$  with batch size 64. We reduce the learning rate by a factor of 0.7 if the BLEU-4 score does not increase for 5 epochs. On MeineDGS, the transformer outperforms GPT in terms of BLEU. However, when considering BLEURT, which reflects semantic similarity, GPT achieves better results (Table 3).

On the DGS-20 videos, we evaluate the same Transformer model that was trained on the MeineDGS dataset. ChatGPT significantly outperforms transformers in all metrics (Table 4). This performance can be attributed to these factors: the spotter achieves high accuracy due to the limited vocabulary, and ChatGPT effectively maps detected gloss sequences into spoken language sentences. These results highlight that when gloss detection is reliable, prompting an LLM like ChatGPT can produce high-quality translations without any task-specific fine-tuning.

**Table 3: Performance of our approach on MeineDGS-V test set. Color-coded for easier comparison: GPT vs. Transformer (blue), Spotter vs. Sub-GT (orange).**

Method	B-1	B-2	B-3	B-4	BLEURT
Spotter+GPT	14.82	4.19	1.45	0.64	<b>21.62</b>
Spotter+Transformer	<b>19.5</b>	<b>6.13</b>	<b>2.48</b>	<b>1.08</b>	19.01
Sub-GT+GPT	16.65	6.45	3.02	1.55	29.72

**Table 4: Performance of our approach on DGS-20. Color-coded for easier comparison: GPT vs. Transformer (blue), Spotter vs. GT (orange).**

Model	B-1	B-2	B-3	B-4	BLEURT
Spotter+GPT	<b>38.25</b>	<b>24.13</b>	<b>15.81</b>	<b>9.12</b>	<b>46.93</b>
Spotter+Transformer	18.65	4.91	1.79	0.92	19.85
GT+GPT	68.74	59.23	52.21	46.22	79.04

**Table 5: Qualitative results of the proposed method.**  
**Examples 1-3 are from MeineDGS; Examples 4-6 are from DGS-20.**

GT Glosses	NUR2B NUM-EINER1A:1d PROGRAMM1A ICH1 DDR4 ES-GIBT3
Spoken Language	Es gab nur ein Programm in der DDR. (There was only one program in the GDR.)
Spotter	NUM-EINER PROGRAMM DDR ES-GIBT
Spotter+GPT	Es gibt nur ein Programm in der DDR. (There is only one program in the GDR.)
GT Glosses	SEHR-VIEL2 FEIN1 ESSEN1
Spoken Language	Dort gibt es sehr gutes Essen. (There is very good food there.)
Spotter	SEHR-VIEL GUT ESSEN
Spotter+GPT	Es gibt sehr viel gutes Essen. (There is a lot of good food.)
GT Glosses	LETZTE URLAUB SARDINIEN INDEX
Spoken Language	Letztens war ich im Urlaub auf Sardinien. (I was recently on holiday in Sardinia.)
Spotter	LETZTE MEHR
Spotter+GPT	Letzte Woche gab es mehr Regen als erwartet. (Last week there was more rain than expected.)
GT Glosses	FAMILIE1 ESSEN1 ABEND2 RESTAURANT1
Spoken Language	Die Familie isst abends im Restaurant. (The family eats in the restaurant in the evening.)
Spotter	FAMILIE ESSEN ABEND RESTAURANT
Spotter+GPT	Die Familie geht am Abend ins Restaurant, um zu essen. (The family goes to the restaurant to eat in the evening.)
GT Glosses	DOLMETSCHER1 SPRECHEN3 FAMILIE1 MADCHEN1
Spoken Language	Der Dolmetscher spricht mit der Familie des Mädchens. (The interpreter speaks to the girl's family.)
Spotter	DOLMETSCHER MIT MADCHEN SPRECHEN
Spotter+GPT	Der Dolmetscher spricht mit dem Mädchen. (The interpreter speaks to the girl.)
GT Glosses	DOZENT1 MORGEN1B FAHRRAD1 UNIVERSITAT1
Spoken Language	Der Dozent fährt morgens mit dem Fahrrad zur Universität. (Tomorrow the lecturer will ride his bike to the university.)
Spotter	DOZENT MORGEN FAHRRAD LUFT
Spotter+GPT	Morgen fahre ich mit dem Fahrrad zur Uni, weil mein Dozent sagt, frische Luft ist wichtig. (Tomorrow I'll ride my bike to university because my lecturer says fresh air is important.)

#### 4.4 Qualitative results

We provide qualitative results of our approach in Table 5. When the spotter successfully detects the majority of glosses, GPT effectively generates high-quality spoken sentences from the glosses (Example 1, 2, and 4). Even in cases where the glosses are incomplete or contain additional glosses, ChatGPT still preserves semantic coherence with information gaps or the generation of new content (Examples 5, and 6). On the other hand, not surprisingly, ChatGPT's performance heavily depends on the quality of the input glosses. When the spotter fails to detect glosses, ChatGPT generates incorrect sentences (Example 3) or "No translation".

## 5 Conclusion

In this paper, we proposed a novel sign language translation framework that combines a sign spotter with a Large Language Model (LLM), specifically ChatGPT, to generate spoken language sentences from sign language videos. Our method does not require any end-to-end SLT model training and leverages prompt-based inference for gloss-to-text generation.

Experimental results on MeineDGS-V and a newly collected DGS-20 dataset show that Spotter+GPT produces coherent and

semantically accurate sentences, especially when the spotter successfully identifies relevant glosses. This indicates that leveraging LLMs offers a promising and flexible alternative to traditional gloss-to-text pipelines.

Our system can process video inputs from both pre-recorded datasets and live capture devices such as webcams. This flexibility allows us to support potential real-time applications.

However, our system's performance is inherently constrained by the vocabulary and accuracy of the sign spotter. When the spotter fails to detect critical glosses, the translation quality drops significantly. This highlights the importance of high-quality gloss spotting. Nevertheless, our approach can be adapted to specialized sign language interpretation by fine-tuning the spotter on domain-specific gloss sets. A future direction may include expanding the vocabulary of the spotter to increase the range of recognized glosses.

## Acknowledgments

We would like to thank Necati Cihan Camgoz for the valuable discussions and feedback. This work was supported by the SNSF project 'SMILE II' (CRSII5 193686), the Innosuisse IICT Flagship (PFFS-21-47), EPSRC grant APP24554 (SignGPT-EP/Z535370/1), and through funding from Google.org via the AI for Global Goals scheme. This work reflects only the author's views and the funders are not



responsible for any use that may be made of the information it contains.

## References

- [1] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. 2020. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI* 16. Springer, 35–53.
- [2] Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, et al. 2021. Bbc-oxford british sign language dataset. *arXiv preprint arXiv:2111.03635* (2021).
- [3] Matyáš Bohacek and Marek Hruš. 2023. Learning from What is Already Out There: Few-shot Sign Language Recognition with Online Dictionaries. In *2023 17th International Conf. on Automatic Face and Gesture Recognition (FG)*. IEEE.
- [4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [5] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural Sign Language Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [6] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10023–10033.
- [7] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.
- [8] Zhigang Chen, Benjia Zhou, Jun Li, Jun Wan, Zhen Lei, Ning Jiang, Quan Lu, and Guoqing Zhao. 2024. Factorized learning assisted with large language model for gloss-free sign language translation. *arXiv preprint arXiv:2403.12556* (2024).
- [9] Rungpeng Cui, Hu Liu, and Changshui Zhang. 2019. A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia* 21, 7 (2019), 1880–1891.
- [10] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metzger, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2sign: a large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2735–2744.
- [11] Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. 2024. Llms are good sign language translators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18362–18372.
- [12] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*. 369–376.
- [13] Youngjoon Jang, Haran Raajesh, Liliane Momeni, Gül Varol, and Andrew Zisserman. 2025. Lost in translation, found in context: Sign language translation with contextual cues. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 8742–8752.
- [14] Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. 2021. Skeleton aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3413–3423.
- [15] Hamid Reza Vaezi Joze and Oscar Koller. 2019. Ms-asl: A large-scale data set and benchmark for understanding american sign language. In *In British Machine Vision Conference*.
- [16] Reiner Konrad, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, Oliver Böse, Elena Jahn, and Marc Schulder. 2020. MEINE DGS–annotiert. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release / MY DGS – annotated. Public Corpus of German Sign Language, 3rd release. (2020).
- [17] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 1459–1469.
- [18] Han Liang, Chengyu Huang, Yuecheng Xu, Cheng Tang, Weicai Ye, Juzhe Zhang, Xin Chen, Jingyi Yu, and Lan Xu. 2024. LLaVA-SLT: Visual Language Tuning for Sign Language Translation. *arXiv preprint arXiv:2412.16524* (2024).
- [19] Shushen Manakhimova, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2023. Linguistically Motivated Evaluation of the 2023 State-of-the-art Machine Translation: Can ChatGPT Outperform NMT?. In *Proceedings of the Eighth Conference on Machine Translation*. 224–245.
- [20] Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-Bonet, Roman Grundkiewicz, et al. 2022. Findings of the first wmt shared task on sign language translation (wmt-slt22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*. 744–772.
- [21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.
- [22] Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards Making the Most of ChatGPT for Machine Translation. Available at SSRN 4390455 (2023).
- [23] Matt Post. 2018. A call for clarity in reporting BLEU scores. *arXiv preprint arXiv:1804.08771* (2018).
- [24] Amy Pu, Hyung Won Chung, Ankur P Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of EMNLP*.
- [25] Junfu Pu, Wengang Zhou, and Houqiang Li. 2019. Iterative alignment network for continuous sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4165–4174.
- [26] Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*, Accepted by *Empirical Methods in Natural Language Processing (EMNLP) 2023*. (2023).
- [27] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).
- [28] Prajit Ramachandran, Barret Zoph, and Quoc V Le. 2017. Searching for activation functions. *arXiv preprint arXiv:1710.05941* (2017).
- [29] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2022. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5141–5151.
- [30] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. BLEURT: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696* (2020).
- [31] Nada Shahin and Leila Ismail. 2023. ChatGPT, Let Us Chat Sign Language: Experiments, Architectural Elements, Challenges and Research Directions. In *2023 International Symposium on Networks, Computers and Communications*. IEEE.
- [32] Ozge Mercanoglu Sincan, Necati Cihan Camgoz, and Richard Bowden. 2023. Is context all you need? scaling neural sign language translation to large domains of discourse. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1955–1965.
- [33] Ozge Mercanoglu Sincan and Hacer Yalim Keles. 2020. Autsl: A large scale multimodal turkish sign language dataset and baseline methods. *IEEE Access* 8 (2020), 181340–181355.
- [34] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. 2013. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*. PMLR, 1139–1147.
- [35] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [37] Manuel Vázquez-Enríquez, Jose L Alba-Castro, Laura Docío-Fernández, and Eduardo Rodríguez-Banga. 2021. Isolated sign language recognition with multi-scale spatial-temporal graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3462–3471.
- [38] Fangyun Wei and Yutong Chen. 2023. Improving continuous sign language recognition with cross-lingual signs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 23612–23621.
- [39] Ryan Wong, Necati Cihan Camgöz, and Richard Bowden. 2022. Hierarchical I3D for Sign Spotting. In *European Conference on Computer Vision*. Springer, 243–255.
- [40] Ryan Cameron Wong, Necati Cihan Camgöz, and Richard Bowden. 2024. Sign2GPT: leveraging large language models for gloss-free sign language translation. In *ICLR 2024: International Conference on Learning Representations*.
- [41] Huijie Yao, Wengang Zhou, Hao Feng, Hezhen Hu, Hao Zhou, and Houqiang Li. 2023. Sign language translation with iterative prototype. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15592–15601.
- [42] Biao Zhang, Mathias Müller, and Rico Sennrich. 2022. SLTUNET: A Simple Unified Model for Sign Language Translation. In *The Eleventh International Conference on Learning Representations (ICLR)*.
- [43] Jiangbin Zheng, Yile Wang, Cheng Tan, Siyuan Li, Ge Wang, Jun Xia, Yidong Chen, and Stan Z Li. 2023. Cvt-slr: Contrastive visual-textual transformation for sign language recognition with variational alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23141–23150.
- [44] Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving Sign Language Translation With Monolingual Data by Sign Back-Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1316–1325.