Assessing the Capabilities of Large Language Models in Translating American Sign Language Gloss to English

Jalal Al-Afandi[*], Péter Pócsi[†], Gábor Borbély[†],
Helga M. Szabó[†], Ádám Rák[*†], Zsolt Robotka[*†], András Horváth[*]

alafandi.mohammad.jalal@hallgato.ppke.hu, { rak.adam, horvath.andras}@itk.ppke.hu
[*]Peter Pazmany Catholic University, Faculty of Information Technology and Bionics

{peter.pocsi, gabor, helga, zsolt}@deepsign.ai
[†] DeepSign Technologies Ltd.

Abstract—**In this paper, we investigate the ability of large language models (LLMs) to translate American Sign Language with GLOSS annoation into English without fine-tuning or architectural modifications. Our findings show that pretrained transformers achieve translation quality comparable to human experts. While prompt engineering enhances accuracy for simpler models, it has minimal impact on more advanced ones. Additionally, when generating multiple translation variants, the first response is typically the most accurate, with subsequent outputs declining in quality. These results underscore the strong zero-shot translation capabilities of LLMs and highlight their potential for scalable ASL-GLOSS translation applications.**

***Keywords*-ASL-GLOSS translation; Generative pretrained transformers, large language models**

I. Introduction

Large Language Models (LLMs) have emerged as a transformative force in natural language processing, demonstrating remarkable versatility across various applications, including text generation, summarization, and machine translation. These models, often referred to as foundation models, are trained on vast corpora of text and possess extensive knowledge of human languages. Their ability to generalize across a wide range of tasks has enabled them to achieve impressive performance, even in low-resource language translation tasks. Recent studies have shown that LLMs excel in one-shot and few-shot learning scenarios [1], where only a limited number of examples are available. This makes them particularly suitable for translating languages with scarce training data.

Among the communities that could greatly benefit from these advancements are deaf and hard-of-hearing individuals. Sign languages serve as the primary mode of communication for these communities; however, the automatic translation of sign languages into spoken or written languages remains a significant challenge [2]. Developing effective translation solutions could substantially enhance accessibility and inclusivity, supporting social integration and improving communication opportunities for these individuals.

TABLE I. Example sentences in English and ASL-GLOSS

English sentence	ASL-GLOSS
There are a lot of studies on speech disorders	STUDY ON SPEECH/ORAL fs-DISORDER A-LOT
While I was a graduate student, in a linguistics class, a professor gave a lecture about syntax.	DURING/WHILE IX-1p GRAD STUDENT IN CLASS_2 LONG-AGO LINGUISTICS TEACH+AGENT TEACH+AGENT DIRECT/EXPLAIN fs-SYNTAX
My mother taught my two brothers and me, so it was easier for us to move around.	part:indef MOTHER TEACH IX-1p+ AND TWO BROTHER EASY MOVE part:indef

Automatic translation of sign languages typically follows a two-step pipeline [3][4], although end-to-end approaches have also been explored [5]. The first step involves recognizing and detecting visual symbols associated with sign language gestures. From these visual inputs, a structured intermediate representation can be derived, such as ASL-GLOSS. ASL-GLOSS serves as a symbolic transcription of American Sign Language (ASL) gestures, capturing the essential lexical components of signs while abstracting away certain nonmanual markers, including facial expressions, eye movements, and contextual cues. Although ASL-GLOSS simplifies the representation of sign language, it remains an incomplete encoding of meaning, as it lacks many elements necessary for full semantic understanding, although this limitation also applies to written text. Table I presents examples of English and ASL-GLOSS sentences.

Existing machine translation solutions for sign-to-English or ASL-GLOSS-to-English tasks typically rely on smaller, domain-specific models trained exclusively on sign and gloss-specific datasets [6]. However, these models often struggle with generalization due to their limited exposure to the target output language. We hypothesize that the broad linguistic knowledge embedded in LLMs can mitigate this issue by providing improved translations and using their comprehensive understanding of syntax, semantics, and common expressions in the target language.

In this work, we explore the capabilities of current LLMs in translating ASL-GLOSS into English. Our objective is to assess the direct translation quality of LLMs on ASL-GLOSS inputs and to establish a baseline accuracy for LLM-based gloss translation. Additionally, by analyzing potential translation errors and corrections, we aim to provide insights into the viability of LLMs as robust components in future sign language translation pipelines.

The remainder of this paper is organized as follows. Section II. provides an overview of the theoretical background and related work relevant to our study. In Section III., we introduce the proposed methodology, including dataset descriptions and evaluation metrics. The results and their analysis are discussed in Section IV., highlighting both quantitative and qualitative findings. Finally, Section V. concludes the paper by summarizing the main contributions and key findings along with the potential avenues for future research.

II. Exisiting solutions

Machine Translation (MT) of ASL encompasses various approaches, each leveraging different technologies to facilitate translation between ASL and spoken or written languages. Key methodologies include:

1) Rule-Based Systems: Early MT systems for ASL utilized rule-based approaches, where linguistic experts encoded grammatical and syntactic rules to map English text to ASL structures [7]. An example is the TEAM prototype, which analyzed English text's syntactic and morphological aspects before accessing a sign synthesizer to produce corresponding ASL signs via a computer-generated human avatar [8].
2) Statistical Machine Translation (SMT): SMT approaches rely on statistical models derived from bilingual corpora to predict translation probabilities. However, the scarcity of large-scale parallel ASL-English corpora has limited the effectiveness of SMT in ASL translation [9].
3) Neural Machine Translation (NMT): Recent advancements in NMT have shown promise in translating spoken languages. Applying NMT to ASL involves training deep learning models on annotated sign language datasets to capture the nuances of ASL grammar and expressions. Challenges include the need for extensive datasets and the complexity of modeling sign language's spatial and temporal aspects [6].
4) Vision-Based Recognition Systems: These systems employ computer vision techniques to interpret sign language from video input [10]. For instance, the Kinect Sign Language Translator utilizes Microsoft's Kinect sensor to capture signers' movements and translate them into spoken language using machine learning and pattern recognition [11].
5) Sensor-Based Recognition Systems: Some approaches use wearable sensors to detect hand movements and positions. For example, SignAloud incorporates gloves equipped with sensors that transliterate ASL into English by tracking hand movements and sending data to a computer system for analysis and translation [12].
6) Hybrid Systems: Combining multiple methodologies, hybrid systems aim to enhance translation accuracy. SignAll integrates computer vision and natural language processing to recognize hand shapes and movements, converting this data into simple English phrases to facilitate real-time ASL translation [13].

Despite these advancements, challenges persist, particularly in accurately interpreting the diverse and complex structures of ASL. Ongoing research aims to address these issues by developing more robust models and incorporating larger, more diverse datasets to improve the reliability and inclusivity of ASL machine translation systems.

III. Methodology

To thoroughly evaluate ASL-GLOSS to English translation, it is essential to carefully consider the data sources and models used in this study. The methodology section outlines our approach to selecting appropriate datasets, choosing relevant language models, and establishing a rigorous evaluation framework. These choices form the foundation for robust and reproducible experimental results.

A. Datasets

To evaluate the performance of LLMs in ASLGLOSS-to-English translation, we conducted an extensive review of available datasets. Our primary objective was to select a dataset that meets several critical criteria. The ideal dataset would be large-scale, contain video recordings of the signing person, provide gloss annotations of the signed sentences, and include high-quality English translations. Video recordings are particularly important as they serve as the most accurate reference for human translations, capturing the full range of visual cues necessary for understanding sign language, including hand movements, facial expressions, and other nonmanual markers. Additionally, we prioritized datasets that feature complex sentence structures and a broad spectrum of topics, ensuring comprehensive coverage of real-world communication scenarios.

However, only a limited number of datasets meet these demanding requirements. The datasets we investigated include:

- English-ASL Gloss Parallel Corpus 2012 (ASLG-PC12): A dataset mapping ASL gloss to formal English text[14]
- American Sign Language Linguistic Research Project (ASLLRP) Data Access Interface (DAI): Contains video recordings with corresponding gloss annotations [15].
- MS-ASL Dataset: A large-scale dataset for isolated sign recognition[16].
- DAI - ASLLVD: A video dataset with ASL lexical items[17].
- ASL Finger Spelling Dataset: Focused on finger-spelling gestures[18].
- WLASL: A large-scale dataset for word-level American Sign Language recognition.[19]
- American Sign Language Lexicon Video Dataset: A comprehensive dataset with video recordings, gloss annotations, and English translations[20].

Among these datasets, the American Sign Language Lexicon Video Dataset proved to be the most suitable for our experiments, as it met all the aforementioned selection criteria. Its combination of video input, detailed gloss annotations, and high-quality English translations makes it an ideal resource for training and evaluating ASL-GLOSS-to-English translation models. Consequently, our experimental work primarily focuses on this dataset.

B. Large Language models

In our investigation, we selected a diverse range of language models to evaluate their performance on the ASL-GLOSS-to-English translation task. Given the rapid advancements in the field, with new models emerging regularly, compiling an exhaustive list is not feasible. However, our selection was guided by several key considerations to ensure a representative and comprehensive assessment.

The selected models fall into two broad categories:

- Large-Scale Proprietary Models: This category includes cutting-edge models such as Claude and ChatGPT, which are accessible exclusively through API-based interfaces. These models are considered among the most complex and sophisticated LLMs available, and we anticipated that their extensive training data and advanced architectures would yield the highest translation accuracy. Despite their closed-source nature, their performance serves as an upper-bound benchmark for comparison.
- Open-Source Models: We also included open-source models, such as LLaMA and DeepSeek. Although these models are typically less complex than their proprietary counterparts, their publicly available architectures and weights offer several advantages. Running these models on-premise enables greater control over execution environments, facilitates further optimization, and allows fine-tuning on domain-specific data. This flexibility is particularly valuable for tailoring models to the nuances of ASLGLOSS translation.

By evaluating models from both categories, we aim to balance performance, transparency, and practical deployability in our study. This comprehensive selection will provide insights into the trade-offs between accuracy and customizability, helping to identify the most suitable models for real-world sign language translation applications.

For the sake of reproducibility, all our experiments, including the code and detailed parameter setups, are available at the following GitHub link to ensure reproducibility: https://github.com/horan85/ASLGloss

IV. Results

Before presenting the experimental results, we summarize the comparative evaluation of various state-of-the-art language models in the ASL-GLOSS to English translation task. This analysis focuses on assessing how model architecture and prompting strategies influence translation quality. Our goal is to understand not only the absolute performance of these models but also how additional linguistic context affects their translation capabilities.

A. Model Comparisons

To systematically assess the performance of our translation models, we curated an evaluation dataset comprising 2,040 ASL-GLOSS-English sentence pairs sourced from the American Sign Language Lexicon Video Dataset. This dataset serves as a benchmark for measuring translation quality and the generalization capabilities of our models.

We conducted experiments with various models under two distinct prompting strategies. In the first setup, models received only a direct translation prompt, instructing them to generate an English sentence from a given ASL-GLOSS input. In the second setup, we supplemented the prompt with a brief explanation of the GLOSS structure (3,000 words in length) and a carefully selected set of twenty example translations to provide additional context and guidance.

As evaluation metrics, we selected the BLEU score and cosine similarity between the embedded representations of the translated and ground-truth sentences. For sentence embeddings, we utilized the CLIP-ViT-B/32 transformer model [21].

Our results for the models without additional descriptions are presented in Table II, which reports the mean performance along with the corresponding variances. To provide a more comprehensive view of the distribution, Figures 1 and 2 illustrate the detailed distributions of BLEU scores and cosine similarities, respectively. These visualizations offer deeper insights into the variability and consistency of model outputs across different evaluation metrics.

Our findings indicate that for more advanced and complex models, such as ChatGPT, Claude, and DeepSeek, the inclusion of structural information and example translations had minimal impact on overall translation quality. This suggests that these models inherently possess a strong ability to interpret ASL-GLOSS sequences and generate fluent English translations, even without explicit guidance on the source language structure.

In contrast, the LLaMA and ChatGPT-mini models exhibited moderate improvements, with increases of approximately 0.03 and 0.04 in cosine similarity and BLEU scores, respectively. However, further investigation is needed to determine whether this robustness extends to less frequent linguistic structures or more complex GLOSS annotations.

B. Model Consistency

Since LLMs generate probabilistic outputs, translation quality can vary due to multiple factors. Additionally, ASL-GLOSS sentences, when extracted without broader context, may have multiple valid interpretations. To assess whether generating multiple translation variants improves accuracy, we examined the effect of allowing GPT-based models to produce several alternative translations for each input.

While output variability can be adjusted by tuning the model's temperature parameter, we did not optimize this aspect. Instead, we instructed the models to generate five

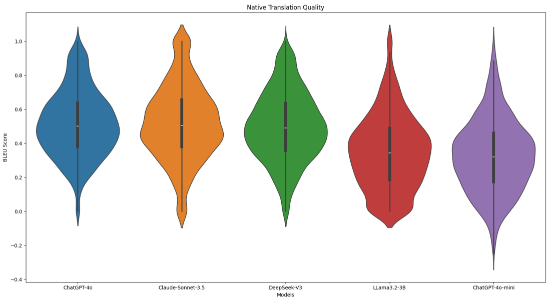

Figure 1. This figure depicts the Blue scores on the American Sign Language Lexicon Video Dataset using various LLM models as a GLOSS to Enlish translation task.

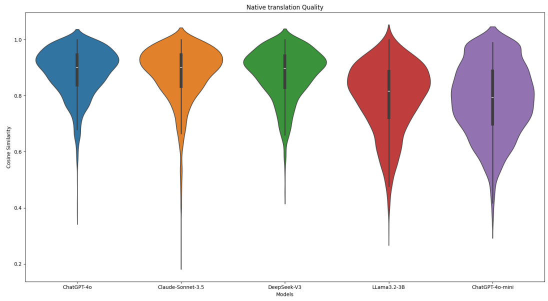

Figure 2. This figure depicts the Cosine similarity values on the American Sign Language Lexicon Video Dataset using various LLM models as a GLOSS to Enlish translation task.

TABLE II. COSINE SIMILARITIES AND BLEU SCORES WITH AND WITHOUT GLOSS DESCRIPTIONS

Model	**Cosine Similarity**	**BLEU Score**
ChatGPT-4o	0.881 ± 0.83	0.514 ± 0.192
ChatGPT-4o with GLOSS description	0.880 ± 0.87	0.512 ± 0.201
Claude Sonnet	0.879 ± 0.94	0.518 ± 0.219
Claude Sonnet with GLOSS description	0.880 ± 0.99	0.518 ± 0.244
DeepSeek V3	0.876 ± 0.83	0.496 ± 0.217
DeepSeek V3 with GLOSS description	0.876 ± 0.88	0.495 ± 0.227
Llama 3.2	0.793 ± 1.23	0.349 ± 0.203
Llama 3.2 with GLOSS description	0.824 ± 1.43	0.374 ± 0.486
ChatGPT-4o-mini	0.787 ± 1.26	0.324 ± 0.213
ChatGPT-4o-mini with GLOSS description	0.814 ± 1.67	0.365 ± 0.455

translation variants per input to evaluate whether this approach enhances translation quality.

Using the same dataset of 2,040 sentences, we selected the two best-performing models (ChatGPT and Sonnet) and tasked them with generating five distinct translations for each ASL-GLOSS input without providing detailed GLOSS descriptions. The BLEU scores for these translations are shown in Figure 3. Similar trends were observed in cosine similarity measurements, though these results are omitted due to space constraints.

Notably, our findings indicate that the first generated translation was consistently the most accurate. As the ranking progressed, translation quality gradually declined, though the differences were minor. This suggests that while generating multiple outputs introduces slight variations, the first translation is generally the most reliable.

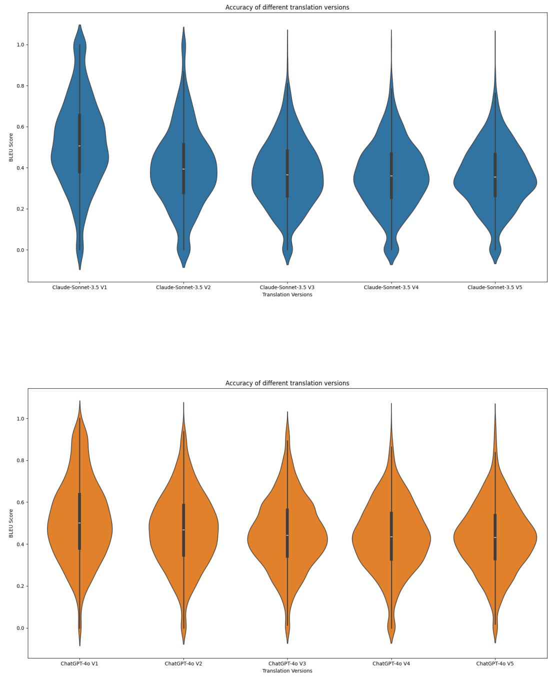

Figure 3. Translation quality (in terms of BLEU scores) when we asked the model to provide multiple variants for Claude-Sonnet (above) and Chat-GPT-4o (below)

-

TABLE III. COSINE SIMILARITIES AND BLEU SCORES ON THE REDUCED DATASET

Model	**Cosine Similarity**	**BLEU Score**
Translator	0.903 ± 0.361	0.582 ± 0.253
ChatGPT-4o	0.893 ± 0.49	0.548 ± 0.163
Claude Sonnet	0.901 ± 0.47	0.560 ± 0.182
DeepSeek V3	0.884 ± 0.78	0.523 ± 0.316
Llama 3.2	0.810 ± 1.23	0.377 ± 0.250

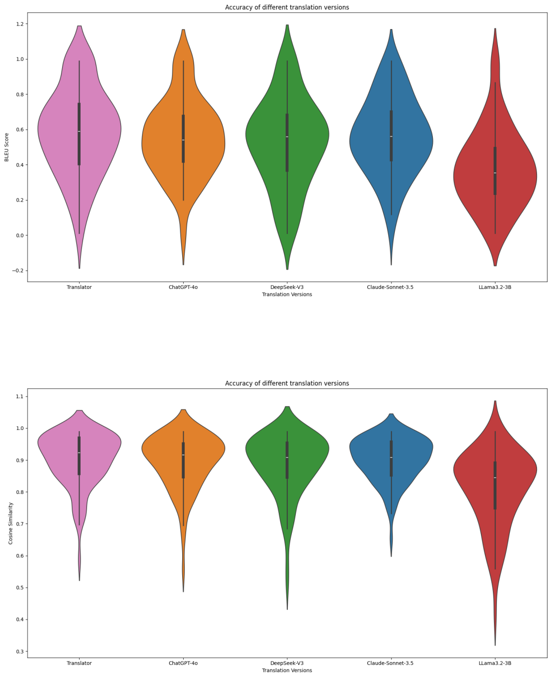

Figure 4. Cosine Similarities (above) and BLEU scores (below) on the 88-sentence dataset comparing a human translator's performance with various LLMs

V. Conclusion and Furute Work

In this paper, we demonstrated the capability of large language models (LLMs) to translate ASL-GLOSS to English without fine-tuning or architectural modifications. Our findings suggest that general-purpose pretrained transformers are viable for this task, achieving translation quality comparable to that of human experts.

A. Key Findings

- Zero-shot translation effectiveness: General-pretrained transformers can effectively translate ASL-GLOSS without additional fine-tuning, highlighting the strong zero-shot capabilities of modern LLMs in handling structured linguistic inputs like GLOSS.
- Limited impact of prompt engineering: While prompt engineering improves translation accuracy for simpler models, it has a negligible effect on more advanced LLMs. This suggests that state-of-the-art models already possess a robust understanding of GLOSS structures without explicit prompting strategies.
- Quality decline in multiple outputs: When LLMs were prompted to generate multiple translation variants, the first response was typically the most accurate, with subsequent translations exhibiting a gradual decline in quality. This suggests that probabilistic generation may introduce increasing errors when multiple outputs are requested.
- Near-human translation accuracy: LLMs achieve translation accuracy close to that of human experts. This underscores their potential to assist or even replace human translators in certain ASL-GLOSS translation tasks, improving scalability and accessibility.

While our results are promising, further research is needed to assess the robustness of LLM-based ASL-GLOSS translation across diverse linguistic structures and complex annotations. Future work could explore fine-tuning approaches, domain adaptation techniques, and real-world deployment scenarios to enhance translation reliability and applicability.

Acknowledgement

The support of the Alfréd Rényi Institute of Mathematics and the following grants: 2018-1.2.1-NKP00008: Exploring the Mathematical Foundations of Artificial Intelligence and TKP2021_02-NVA-27 – Thematic Excellence Program and also 2023-1.1.2-GYORSÍTÓSÁV-2024-00019 of the Hungarian National Research, Development and Innovation Office are gratefully acknowledged.

References

[1] X. Zhang, N. Rajabi, K. Duh, and P. Koehn, "Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with qlora", in *Proceedings of the Eighth Conference on Machine Translation*, 2023, pp. 468–481.

[2] U. Farooq, M. S. M. Rahim, N. Sabir, A. Hussain, and A. Abid, "Advances in machine translation for sign language: Approaches, limitations, and challenges", *Neural Computing and Applications*, vol. 33, no. 21, pp. 14 357–14 399, 2021.

[3] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7784–7793.

[4] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation", in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[5] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 023–10 033.

[6] P. Fayyazsanavi, A. Anastasopoulos, and J. Košecká, "Gloss2text: Sign language gloss translation using llms and semantically aware label smoothing", *arXiv preprint arXiv:2407.01394*, 2024.

[7] J. Porta, F. López-Colino, J. Tejedor, and J. Colás, "A rule-based translation from written spanish to spanish sign language glosses", *Computer Speech & Language*, vol. 28, no. 3, pp. 788–811, 2014.

[8] B. David and P. Bouillon, "Prototype of automatic translation to the sign language of french-speaking belgium evaluation by the deaf community", *Modelling, Measurement and Control C*, vol. 79, no. 4, pp. 162–167, 2018.

[9] A. Othman and M. Jemni, "Statistical sign language machine translation: From english written text to american sign language gloss", *arXiv preprint arXiv:1112.0168*, 2011.

[10] Y. Madhuri, G. Anitha, and M. Anburajan, "Vision-based sign language translation device", in *2013 International Conference on Information Communication and Embedded Systems (ICICES)*, IEEE, 2013, pp. 565–568.

[11] M. Ahmed, M. Idrees, Z. ul Abideen, R. Mumtaz, and S. Khalique, "Deaf talk using 3d animated sign language: A sign language interpreter using microsoft's kinect v2", in *2016 SAI Computing Conference (SAI)*, IEEE, 2016, pp. 330–335.

[12] S. B. Rizwan, M. S. Z. Khan, and M. Imran, “American sign language translation via smart wearable glove technology”, in *2019 International Symposium on Recent Advances in Electrical Engineering (RAEE)*, IEEE, vol. 4, 2019, pp. 1–6.

[13] L. Leeson and H. Haaris, “Signall: A european partnership approach to deaf studies via new technologies”, in *INTED2009 Proceedings*, IATED, 2009, pp. 1270–1279.

[14] A. Othman and M. Jemni, “English-asl gloss parallel corpus 2012: Aslg-pc12”, in *Sign-lang@ LREC 2012*, European Language Resources Association (ELRA), 2012, pp. 151–154.

[15] C. Neidle and S. Sclaroff, “American sign language linguistic research project”, Report, Tech. Rep., 1998.

[16] H. R. V. Joze and O. Koller, “Ms-asl: A large-scale data set and benchmark for understanding american sign language”, *arXiv preprint arXiv:1812.01053*, 2018.

[17] C. Neidle, A. Thangali, and S. Sclaroff, “Challenges in development of the american sign language lexicon video dataset (asllvd) corpus”, in *Proc. 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, Language Resources and Evaluation Conference (LREC) 2012*, 2012.

[18] N. Pugeault and R. Bowden, “Spelling it out: Real-time asl fingerspelling recognition”, in *2011 IEEE International conference on computer vision workshops (ICCV workshops)*, Ieee, 2011, pp. 1114–1119.

[19] D. Li, C. Rodriguez, X. Yu, and H. Li, “Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison”, in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2020, pp. 1459–1469.

[20] V. Athitsos *et al.*, “The american sign language lexicon video dataset”, in *2008 IEEE computer society conference on computer vision and pattern recognition workshops*, IEEE, 2008, pp. 1–8.

[21] A. Radford *et al.*, “Learning transferable visual models from natural language supervision”, in *International conference on machine learning*, PmLR, 2021, pp. 8748–8763.