Augmenting Sign Language Translation Datasets with Large Language Models

Pedro Dal Bianco
III-LIDI
Universidad Nacional de La Plata
pdalbianco@lidi.info.unlp.edu.ar

Jean Paul Nunes Reinhold
CDTEC, Federal University of Pelotas
jean.pnr@inf.ufpel.edu.br

Facundo Quiroga
III-LIDI
Comisión de Investigaciones Científicas
Universidad Nacional de La Plata
fquiroga@lidi.info.unlp.edu.ar

Franco Ronchetti
III-LIDI
Comisión de Investigaciones Científicas
Universidad Nacional de La Plata
fronchetti@lidi.info.unlp.edu.ar

Abstract

Sign language translation (SLT) is a challenging task due to the scarcity of labeled data and the heavy-tailed distribution of sign language vocabularies. In this paper, we explore a novel data augmentation approach for SLT: using a large language model (LLM) to generate paraphrases of the target language sentences in the training data. We experiment with a Transformer-based SLT model (Signformer) on three datasets spanning German, Greek, and Argentinian Sign Languages. For models trained with augmentation, we adopt a two-stage regime: pre-train on the LLM-augmented corpus and then fine-tune on the original, non-augmented training set. Our augmented training sets, expanded with GPT-4-generated paraphrases, yield mixed results. On a medium-scale German SL corpus (PHOENIX14T), LLM augmentation improves BLEU-4 from 9.56 to 10.33. In contrast, a small-vocabulary Greek SL dataset with a near-perfect baseline (94.38 BLEU) sees a slight drop to 92.22 BLEU, and a complex Argentinian SL corpus with a long-tail vocabulary distribution remains around 1.2 BLEU despite augmentation. We analyze these outcomes in relation to each dataset's complexity and token frequency distribution, finding that LLM-based augmentation is more beneficial when the dataset contains a richer vocabulary and many infrequent tokens. To our knowledge, this work is the first to apply LLM paraphrasing to SLT, and we discuss these results with respect to prior data augmentation efforts in sign language translation.

1 Introduction

Sign Language Translation (SLT) aims to convert sign language video into spoken language text, bridging communication between deaf signers and hearing people. It is a multimodal task at the intersection of computer vision and natural language processing, and has seen steady progress in recent years (Camgoz et al., 2018, 2020). However, SLT remains extremely challenging due to the scarcity of large-scale parallel sign-video-to-text datasets (Bragg et al., 2019). Datasets that do exist tend to be limited in domain and have a heavy-tailed vocabulary distribution, with many words appearing only a few times (or even once) in the corpus. For example, the popular RWTH-PHOENIX-Weather 2014T (Phoenix14T) German SL dataset (Camgoz et al., 2018) has a relatively small vocabulary (under 3k words) and a high mean word frequency, making it easier for models to achieve relatively good BLEU scores compared to broader-domain corpora. In contrast, newer, more diverse SLT datasets feature much larger vocabularies and a majority of low-frequency tokens, resulting in very low baseline translation performance. The combination of sparsity and long-tail token distribution poses a major hurdle for training effective SLT models. A quantitative summary of these differences including vocabulary size and the proportion of singletons that drive long-tail effects is provided in Table 2.

Data augmentation is a common strategy to address low-resource settings. In spoken language machine translation, methods like back-translation

Proceedings of the Workshop on Sign Language Processing (WSLP), pages 24–30

and paraphrasing are commonly used to boost performance in low-resource scenarios (Sennrich et al., 2016; Hu et al., 2021). In the context of sign language, prior work has explored various augmentation techniques. Moryossef et al. (2021) generate synthetic gloss–text pairs from monolingual spoken text and report *relative* gains of +19.7% BLEU on NCSLGR (Neidle and Vogler, 2012) and +10.4% on PHOENIX14T (Camgoz et al., 2018). More recently, (Walsh et al., 2025) leveraged Sign Language Production models to generate new sign video samples (either via skeletal pose manipulation or video GANs), yielding up to 19% relative improvement in BLEU score. These approaches augment data on the sign language input (either at the gloss or video level). By contrast, our focus is on augmenting the *text output* of the training pairs using modern LLMs.

Large language models have demonstrated remarkable capabilities in producing paraphrases and diversifying text while preserving meaning. We investigate whether an LLM (GPT-4 in our case) can be used to automatically create multiple paraphrased translations for each sign video, thereby enlarging the effective training set and exposing the translation model to a richer variety of expressions. Our hypothesis is that this can alleviate the impact of rare words and rigid sentence patterns in SLT training data. To our knowledge, this idea has not been explored in prior SLT research, although LLMs have been integrated into SLT pipelines in other ways (e.g., using pretrained text models for the translation decoder (Wong et al., 2024)).

We conduct experiments on three datasets covering different sign languages and levels of complexity: (1) Phoenix14T (German Sign Language) (Camgoz et al., 2018), a weather forecast domain corpus; (2) a Greek Sign Language (GSL) corpus of educational video translations (Voskou et al., 2023); and (3) an Argentinian Sign Language (LSA) corpus derived from the LSA-T dataset (Bianco et al., 2022). We augment each training set by generating three paraphrases per original sentence using GPT-4 (with prompts instructing the model to preserve semantics and most words while varying word order). For augmented models, we first train on the augmented corpus and then fine-tune on the original sentences only. We employ a Transformer translation model based on the Signformer architecture (Yang, 2024). We compare our augmented models against baselines trained solely on the original data.

Our main contributions can be summarized as follows: (1) We introduce LLM-based target-output paraphrasing as a data augmentation technique for sign language translation and release four augmented versions of SLT datasets (covering DGS, GSL, LSA and ISL). (2) We present an empirical evaluation of this augmentation across datasets of varying vocabulary size and complexity, showing that its impact differs markedly: from a modest BLEU-4 improvement in one case to negligible or even slight negative effects in others. We analyze these outcomes and provide hypotheses linking them to dataset properties such as vocabulary breadth and frequency of singletons. All of our code and datasets are publicly available [1].

2 Related Work

Sign Language Translation. Early SLT systems followed a two-stage approach: first performing continuous sign language recognition to predict an intermediate gloss sequence, then translating glosses to text (Camgoz et al., 2018). Glosses are textual labels (often one per sign) that approximate the signed content. While glosses simplify the translation problem, creating gloss annotations is labor-intensive and glosses cannot capture all nuances (facial expressions, classifier constructions, etc.). To avoid these limitations, recent research has shifted toward *gloss-free* SLT, building end-to-end models that map video directly to spoken language text (Camgoz et al., 2020; Chen et al., 2022). Gloss-free SLT is considerably more challenging, typically yielding lower accuracy than gloss-based methods, but it is more scalable since it requires only video-text pairs. Modern gloss-free approaches often employ transformer architectures and have begun incorporating large pretrained models. For example, the Sign2GPT system (Wong et al., 2024) uses a pretrained CLIP visual encoder and a GPT-style language model for decoding, with lightweight adapters, achieving state-of-the-art results on Phoenix14T and CSL-Daily (Chinese Sign Language). (Yang, 2024) introduced *Signformer*, a transformer that eschews any pretrained components and is extremely lightweight (0.57M parameters for a smaller variant), yet it reached competitive performance (second place on Phoenix14T gloss-free leaderboard). Our work builds on a Signformer-like architecture, but using a sequence of body pose keypoints as input.

[1] Url anonimized for review purposes.

Data Augmentation in SLT. The scarcity of sign-to-text data has motivated various augmentation strategies. Aside from simple video augmentations (e.g., mirroring, spatial jitter) commonly used in sign recognition, researchers have proposed more complex methods for SLT. On the sign input level, one approach is to generate synthetic training examples using sign language production models. (Stoll et al., 2020) and others have developed techniques to create sign animations or videos from text; however, the visual quality and realism of generated signs can be limiting. Recent work by (Walsh et al., 2025) took a step forward by employing (i) skeleton-based motion synthesis and stitching, and (ii) generative adversarial models (SignGAN, SignSplat) to produce artificial sign video variations, yielding *relative* improvements in BLEU of up to ∼19% on benchmark SLT datasets. Complementarily, in *sign language recognition* (SLR), dynamic sign generation has also proven effective: works like Rios et al. (2025) introduce *HandCraft*, a lightweight generator that produces synthetic sign sequences and, through synthetic-data pretraining, establishes new state-of-the-art results on LSFB and DiSPLaY—further supporting the value of sign-level augmentation for recognition. On the text output, data augmentation is less explored in SLT. (Moryossef et al., 2021) augmented the text output of a gloss-to-text translator by creating paraphrase pairs from monolingual data with heuristic rules, effectively expanding data via pseudo-gloss generation. In broader NLP, LLMs like GPT-3/4 have been used to generate paraphrases or new training samples for low-resource tasks (Davoodi et al., 2022). In this work, we apply a similar idea specifically to SLT: using an LLM to rephrase ground-truth translations in order to introduce lexical and syntactic variety. This approach does not require any additional sign data and thus is complementary to sign-level augmentation methods. We compare our results with prior augmentation approaches and discuss scenarios where text augmentation might be preferable or vice versa.

3 Methodology

3.1 Model Architecture

Our baseline model is inspired by Signformer (Yang, 2024), a recent transformer-based SLT model designed for efficiency. We adopt a simplified version of Signformer in which, instead of feeding in spatio-temporal visual embeddings (e.g., CNN features from video frames), we use pose keypoints extracted from each video frame. Specifically, we utilize the MediaPipe Holistic (Maia et al., 2025) model to obtain 2D coordinates of the signer's body, hands, and face key landmarks for each frame. These pose landmarks (in total, we use 33 body pose points, 21 points for each hand, and a subset of facial landmarks relevant to mouth and eyebrows) are concatenated into a feature vector per frame, yielding a time-series of pose features. We then project this pose feature vector into the model's embedding space via a linear layer. This serves as the input to the encoder. By using skeleton data, we drastically reduce the input dimensionality and remove background noise, potentially enabling faster training and inference suitable for edge devices (Yang, 2024). However, this comes at the cost of losing some visual information (like detailed appearance, color, or subtle gestures not captured by keypoints). Prior findings suggest pose-based approaches may slightly lag behind image-based models in translation quality, especially on unconstrained content (Zelezný et al., 2025). We acknowledge this trade-off; indeed, our model's absolute BLEU scores are lower than state-of-the-art results that use full video frames (see Section 5). Nonetheless, the *relative* comparisons (with vs. without augmentation) remain meaningful within our setup.

3.2 LLM-Based Data Augmentation

To augment the training data, we employ GPT-4 as a paraphrase generator. For each video-sentence pair (V, T) in the original training set (where T is the ground-truth spoken language translation of the sign video V), we generate $N = 3$ additional sentences T'_1, T'_2, T'_3 that convey the same meaning as T. We design a prompt to guide GPT-4 to produce high-quality paraphrases that preserve semantics and key vocabulary. The prompt (shown in figure 1) attempts to generate paraphrases that are close to the original sentence in vocabulary and style, while introducing some variation (particularly in word order and occasional synonyms). The constraint to reuse 70% of words is intended to prevent GPT-4 from rephrasing too freely and possibly introducing unfamiliar vocabulary that might confuse the translation model. We adjusted the prompt for each target language (e.g., Spanish for LSA, Greek for GSL, etc.) accordingly.

For each original sign video V, we thus obtain 3

translations: the original T and three paraphrases T'_1, T'_2, T'_3. During training we materialize these as 4 separate examples (V, T), (V, T'_1), (V, T'_2), (V, T'_3) (i.e., V is repeated four times with each textual variant). Figure 1 summarizes the overall augmentation pipeline. As concrete illustrations of the augmentation, Table 1 shows three training instances from RWTH-Phoenix datasets and their LLM generated paraphrases.

3.3 Training Schedule

We compare two conditions:

- **Baseline:** train the model on the original (non-augmented) training set only.
- **+Augmentation:**
  - *Stage1:* pre-train on the augmented corpus (original targets + three GPT-4 paraphrases per instance).
  - *Stage2:* fine-tune on the original training set only, to realign the decoder distribution with the reference phrasing and reduce drift toward rare paraphrastic variants. Unless otherwise stated, all hyperparameters are kept identical across conditions; early stopping is performed on the same development set.

4 Datasets and Evaluation

We evaluate our approach on three sign language translation datasets that differ notably in linguistic diversity, recording conditions, and lexical structure—factors that strongly influence how data augmentation behaves.

The **PHOENIX14T** dataset (Camgoz et al., 2018) contains weather broadcast recordings in German Sign Language (DGS) with corresponding German text. It is a real-world corpus characterized by consistent domain-specific phrasing and limited topic variation. Although this repetitiveness simplifies translation, the naturally recorded conditions introduce visual variability across signers and sessions, maintaining a moderate level of linguistic and visual complexity.

In contrast, the **GSL** dataset (Adaloglou et al., 2020) is recorded under controlled laboratory conditions, featuring a small group of signers repeatedly performing a restricted set of predefined sentences. As a result, it exhibits low linguistic and visual variability, with high redundancy across samples and virtually no rare tokens. This simplicity allows models to easily memorize sentence structures and reach near-perfect BLEU scores, but at the cost of generalization.

Finally, the **LSA-T** dataset (Bianco et al., 2022) comprises real-world videos from diverse sources, with a wide range of signers, lighting, and signing styles. Its naturalistic, spontaneous signing and extensive Spanish vocabulary make it a far more challenging dataset. The high proportion of singletons and irregular phrasing create a long-tail distribution, resulting in sparse lexical coverage and low baseline translation accuracy. This makes LSA-T particularly valuable for testing augmentation strategies aimed at mitigating data scarcity and improving robustness under realistic conditions.

Together, these datasets span a spectrum from controlled and repetitive to unconstrained and diverse, providing an ideal testbed for assessing how LLM-based paraphrasing interacts with varying levels of linguistic and visual complexity. Table 2 quantitavely describes mentioned datasets.

For all datasets, we preprocessed the videos with MediaPipe to extract pose sequences, as described above. We then normalized coordinate values and frame rates for input to the model (following steps similar to (Železný et al., 2023)). The text was lowercased and tokenized; we built a separate vocabulary for each language (German, Greek, Spanish) with a size of 5,000 tokens, ensuring coverage of all training words. We evaluate translation quality using case-insensitive BLEU-4 (Papineni et al., 2002) on the test set.

5 Results and Analysis

Table 3 reports BLEU-4 on the test sets for the Baseline vs. the two-stage **+Augmentation** setup.

Overall trends. Phoenix14T shows a small but consistent gain (+0.77 BLEU). Given its moderately rich yet formulaic domain, exposing the decoder to paraphrastic re-orderings appears to improve generalization beyond memorized templates, and the subsequent fine-tuning on original references helps keep the output close to the evaluation style. In contrast, the GSL subset starts with an exceptionally high baseline (94.38 BLEU), indicating substantial overlap and low linguistic variability between training and test. In this near-saturated regime, even with our final fine-tuning stage, augmentation slightly hurts (92.22 BLEU): the decoder learns alternative, semantically valid phrasings that do not exactly match the single reference, and the

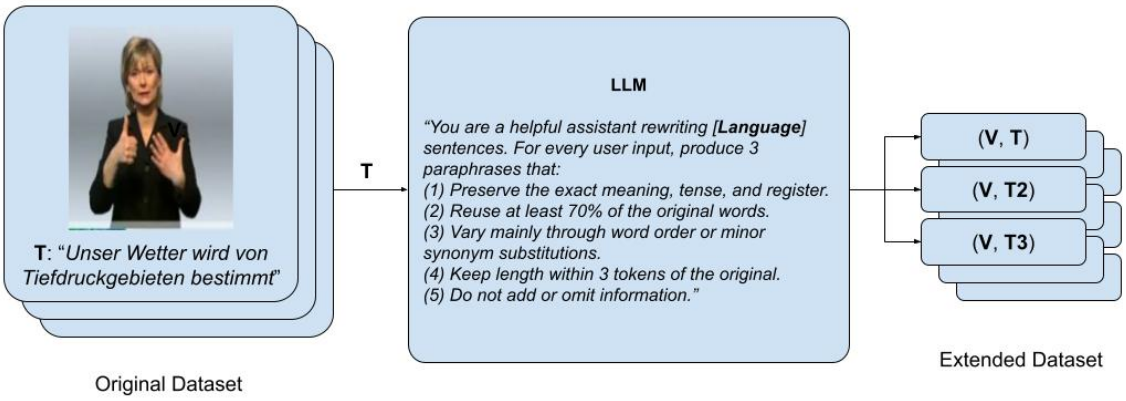

Figure 1: LLM-augmented SLT pipeline. For each video–text pair (V, T) we use an LLM to generate three paraphrases (T'_1, T'_2, T'_3) that preserve meaning while introducing limited lexical/syntactic variety. We then adopt a two-stage regime: (i) pre-train on the augmented corpus (all T and T'_i), (ii) fine-tune on the original targets T only. At test time, the model translates directly from the sign input to text.

Original (reference)	**LLM paraphrases**
tiefdruckgebiete bestimmen unser wetter *low-pressure areas determine our weather*	• Unser Wetter wird von Tiefdruckgebieten bestimmt. *Our weather is determined by low-pressure areas.* • Die Bestimmung unseres Wetters erfolgt durch Tiefdruckgebiete. *The determination of our weather is due to low-pressure areas.*
auch mit den temperaturen geht es aufwärts *the temperatures are also rising*	• Auch die Temperaturen steigen an. *The temperatures are also increasing.* • Die Temperaturen gehen ebenfalls nach oben. *The temperatures are also going up.*
eine gewitterfront überquert deutschland von west nach ost *a thunderstorm front crosses Germany from west to east*	• Eine Gewitterfront zieht von Westen nach Osten über Deutschland. *A thunderstorm front moves from west to east across Germany.* • Von Westen nach Osten überquert eine Gewitterfront Deutschland. *From west to east, a thunderstorm front crosses Germany.*

Table 1: Original training references paired with their GPT-4 paraphrases from the PHOENIX14T dataset.

fine-tune does not fully eliminate these variants. Finally, the reduced LSA-T subset remains extremely low (around 1.2 BLEU) in both settings; paraphrasing largely preserves the same rare content words (by design of our prompt) and thus does not mitigate the core issue: severe data sparsity on the sign inputs and a very heavy-tailed token distribution.

Data characteristics matter. The observed utility of LLM paraphrasing correlates with vocabulary breadth and the prevalence of infrequent tokens. When the dataset offers enough lexical variety (Phoenix14T), paraphrastic exposure helps the model handle word-order and light lexical alternations encountered at test time. When the task is artificially simple (our GSL subset), increased output variety degrades single-reference BLEU despite the final fine-tune. When the vocabulary is extremely sparse (our reduced LSA-T subset), paraphrasing the target alone does not address full coverage: many content signs/words are never learned well enough for the decoder to benefit from the text augmentation.

On pose inputs. Our absolute Phoenix14T scores (around 10 BLEU) are well below SOTA that use full video features and/or gloss supervision (22–24 BLEU). Likely contributors include our small model size, reliance on 2D pose keypoints (which may miss mouthing and subtle facial cues), and the absence of an intermediate gloss stage (Yang, 2024; Maia et al., 2025). Nevertheless, within this consistent pose-based setup, the two-stage augmentation policy yields the relative effects summarized above.

6 Conclusion

We presented a study on augmenting SLT training data by generating paraphrase variations of the target text using an LLM, combined with a two-stage training schedule that pre-trains on augmented text and then fine-tunes on the original data. Across multiple sign languages, this strategy yields a modest improvement on a medium-complexity dataset

Statistic	PHOENIX14T (DGS)	GSL	LSA-T (LSA)
Language (target)	German	Greek	Spanish
Sign language	DGS	GSL	LSA
Real-world footage	Yes	No	Yes
No. of signers	9	7	103
Duration [h]	10.71	9.51	21.78
Samples (clips)	7,096	10,295	8,459
Unique sentences	5,672	331	8,102
% unique sentences	79.93%	3.21%	95.79%
Vocabulary size (types)	2,887	N/A	14,239
Singletons (types with count=1)	1,077	0	7,150
% singletons	37.3%	0%	50.21%
Resolution	210×260	848×480	1920×1080
FPS	25	30	30

Table 2: Corpus statistics for the three datasets used in our experiments. The bottom block highlights lexical properties related to long-tail behavior (vocabulary size and proportion of singletons).

Dataset	Baseline (BLEU-4)	+Augmentation (BLEU-4)
PHOENIX14T (DGS)	9.56	10.33
GSL (Greek)	94.38	92.22
LSA (Spanish)	1.18	1.19

Table 3: Test BLEU-4 for baseline vs. LLM-augmented training on three datasets.

(Phoenix14T), but negligible or negative effects on extremely simple (GSL subset) or extremely sparse (reduced LSA-T subset) settings. These results suggest that LLM-based target output augmentation is not a one-size-fits-all solution; its usefulness depends on properties like vocabulary diversity and data sufficiency.

In addition, we demonstrated a pose-based SLT modeling approach that, while not achieving SOTA accuracy, allowed us to efficiently experiment with data augmentation. An interesting avenue for future work is to combine sign level and output text augmentation: e.g., use sign synthesis to generate new training signs for existing sentences, and simultaneously use text paraphrasing to generate new sentences for existing signs. Such a combination could address both the lack of visual-sign variations and the lack of linguistic variations. Another direction is to apply our augmentation in a scenario with multiple reference translations for evaluation; we hypothesize this would show clearer gains of the method, as single-reference BLEU can penalize legitimate paraphrases even after fine-tuning.

Finally, while we used a powerful proprietary LLM (GPT-4) to generate our paraphrases, it would be valuable to investigate if similar benefits can be obtained with open-source LLMs or simpler neural paraphrasers, and test different variations of the prompt, which would make this approach more accessible and reproducible for the research community.

References

Nikolas Adaloglou, Theocharis Chatzis, Ilias Papastratis, Andreas Stergioulas, Georgios Th Papadopoulos, Vassia Zacharopoulou, George J Xydopoulos, Klimnis Atzakas, Dimitris Papazachariou, and Petros Daras. 2020. A comprehensive study on sign language recognition methods. *arXiv preprint arXiv:2007.12530*.

Pedro Dal Bianco, Gast'on R'ıos, Franco Ronchetti, Facundo Quiroga, Oscar Stanchi, Waldo Hasperu'e, and Alejandro Rosete. 2022. Lsa-t: The first continuous argentinian sign language dataset for sign language translation. In *Advances in Artificial Intelligence – IBERAMIA 2022*, volume 13788 of *Lecture Notes in Computer Science*, page 293–304. Springer, Cham.

Danielle Bragg, Oscar Koller, Miriam Bellard, Larwan Berke, Naomi Caselli, and etc. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. *ACM Transactions on Accessible Computing*, 12(2):5:1–5:44.

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)*, pages 7784–7793.

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*, pages 10020–10030.

Shizhe Chen, Yuecong Wang, and etc. 2022. Two stream transformer networks for sign language trans-

lation. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ehsan Davoodi and 1 others. 2022. Improving low-resource classification via large language models for data augmentation. In *Proceedings of the 60th Annual Meeting of the ACL (Short Papers)*.

Xinyu Hu and 1 others. 2021. Text data augmentation made simple by leveraging llms: A case study on low-resource nlu tasks. In *Proceedings of the EMNLP 2021 (Findings)*.

Wesley Maia, António M. Lopes, and Sérgio A. David. 2025. Automatic sign language to text translation using mediapipe and transformer architectures. *Neurocomputing*, 642:130421.

Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. Data augmentation for sign language gloss translation. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL) at MTSummit*, pages 1–11, Virtual. Association for Machine Translation in the Americas.

Carol Neidle and Christian Vogler. 2012. A new web interface to facilitate access to corpora: Development of the asllrp data access interface (dai). In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon (LREC 2012)*, Istanbul, Turkey. European Language Resources Association (ELRA).

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*, pages 311–318.

Gaston Gustavo Rios, Pedro Dal Bianco, Franco Ronchetti, Facundo Quiroga, Oscar Stanchi, Santiago Ponte Ahón, and Waldo Hasperué. 2025. Handcraft: Dynamic sign generation for synthetic data augmentation. *arXiv preprint arXiv:2508.14345*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the ACL*.

Stefanie Stoll and 1 others. 2020. Text2sign: Towards sign language production using neural machine translation and generative adversarial networks. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

Andreas Voskou, Konstantinos P. Panousis, Harris Partaourides, Kyriakos Tolias, and Sotirios Chatzis. 2023. A new dataset for end-to-end sign language translation: The greek elementary school dataset. *arXiv preprint arXiv:2310.04753*.

Harry Walsh, Maksym Ivashechkin, and Richard Bowden. 2025. Using sign language production as data augmentation to enhance sign language translation. *arXiv preprint arXiv:2506.09643*.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2024. Sign2gpt: Leveraging large language models for gloss-free sign language translation. In *International Conference on Learning Representations (ICLR)*.

Eta Yang. 2024. Signformer is all you need: Towards edge ai for sign language. *arXiv preprint arXiv:2411.12901*.

Tomáš Železný, Jakub Straka, Václav Javorek, Ondřej Valach, Marek Hrúz, and Ivan Gruber. 2023. Exploring pose-based sign language translation: Ablation studies and attention insights. *arXiv preprint arXiv:2507.01532*.

Tomáš Zelezný, Jakub Straka, Václav Javorek, Ondřej Valach, Marek Hrúz, and Ivan Gruber. 2025. Exploring pose-based sign language translation: Ablation studies and attention insights. *arXiv preprint arXiv:2507.01532*.