Multilingual Sign Language Translation with Unified Datasets and Pose-Based Transformers

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

Oscar Stanchi
CONICET
III-LIDI
ostanchi@lidi.info.unlp.edu.ar

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 languages are highly diverse across countries and regions, yet most Sign Language Translation (SLT) work remains monolingual. We explore a unified, *multi-target* SLT model trained jointly on four sign languages (German, Greek, Argentinian, Indian) using a standardized data layer. Our model operates on pose keypoints extracted with MediaPipe, yielding a lightweight and *dataset-agnostic* representation that is less sensitive to backgrounds, clothing, cameras, or signer identity while retaining motion and configuration cues. On RWTH-PHOENIX-Weather 2014T, Greek Sign Language Dataset, LSA-T, and ISLTranslate, naive joint training under a fully shared parameterization performs worse than monolingual baselines; however, a simple two-stage schedule: multilingual pre-training followed by a short language-specific fine-tuning, recovers and *surpasses* monolingual results on three datasets (PHOENIX14T: $+0.15$ BLEU-4; GSL: $+0.74$; ISL: $+0.10$) while narrowing the gap on the most challenging corpus (LSA-T: -0.24 vs. monolingual). Scores span from BLEU-4≈ 1 on open-domain news (LSA-T) to > 90 on constrained curricula (GSL), highlighting the role of dataset complexity. We release our code to facilitate training and evaluation of multilingual SLT models.

1 Introduction

Sign Language Translation (SLT) aims to convert sign language videos into spoken or written language text, helping bridge communication between deaf and hearing communities. SLT research has concentrated mostly on single-language benchmarks. Most notably, German Sign Language (DGS) with RWTH-PHOENIX-Weather 2014T has typically been used as baseline (Camgoz et al., 2018). Subsequently, transformer-based approaches demonstrated steady improvements (Camgoz et al., 2020), yet the diversity of sign languages and the scarcity of labeled data make it impractical to build and maintain one system per language. In contrast, multilingual modeling has transformed spoken/written machine translation (MT): a single shared model with target-language control tokens can learn to translate among many languages and even generalize in low-resource settings (Johnson et al., 2017). Bringing these ideas into SLT is promising but still relatively new. Recent work has shown the feasibility of multilingual SLT with architectural mechanisms to regulate parameter sharing across languages (Yin et al., 2022), and with clustering strategies to mitigate interference by grouping related languages (Zhang et al., 2025); in parallel, scaling data and directions is beginning to push SLT beyond narrow domains (Zhang et al., 2024). However, evaluation setups differ: some studies prefer many-to-one (many sign languages $\rightarrow$ one spoken language) for comparability, while others explore many-to-many configurations with multiple spoken targets, leaving open how far a *fully shared*, standard architecture can go when each sign language is translated into its *own* spoken language.

We address this question by training a single multilingual SLT model across four sign

Proceedings of the Workshop on Sign Language Processing (WSLP), pages 31–36

languages: DGS in RWTH-PHOENIX-Weather 2014T (DGS→German) (Camgoz et al., 2018), the Greek Sign Language Dataset (GSL→Greek) (Adaloglou et al., 2020), LSA-T (Argentinian Sign Language; LSA→Spanish) (Bianco et al., 2023), and ISLTranslate (Indian Sign Language; ISL→English) (Joshi et al., 2023). In this work we adapt *Signformer* (Yang, 2024) to operate on pose keypoints (hands, body, selected facial landmarks) extracted with MediaPipe (Lugaresi et al., 2019) instead of on CNN-derived visual embeddings. This choice yields a lightweight pipeline and can encourage cross-lingual transfer over motion patterns, albeit at the cost of some visual nuance in fine handshape/face details (for which robustness techniques continue to improve (Moryossef, 2024)). Practically, we unify data preparation across these corpora using an open-source library that standardizes formats and preprocessing, lowering barriers to multilingual experimentation.[1]

Our contributions can be listed as:

- **A multi-target multilingual SLT model** that translates each sign language into its *native spoken language* within a single, fully shared Transformer with no language-specific routing, complementing prior multilingual SLT designs that add sharing controls (Yin et al., 2022; Zhang et al., 2025).
- **A unified, open-source data layer** that harmonizes formats and preprocessing across RWTH-PHOENIX-Weather 2014T, Greek Elementary, LSA-T, and ISLTranslate, enabling streamlined multilingual training and evaluation (Bianco, 2025; Camgoz et al., 2018; Adaloglou et al., 2020; Bianco et al., 2023; Joshi et al., 2023).
- **A pose-keypoint adaptation of Signformer** (Yang, 2024) that replaces frame-based encoders with MediaPipe/BlazePose landmarks (Lugaresi et al., 2019; Bazarevsky et al., 2020), producing an efficient model suitable for cross-lingual sharing and deployment.
- **An empirical study of multilingual transfer** on four typologically and domain-diverse sign languages, showing that multilingual pre-training plus light language-specific fine-tuning *surpasses* monolingual baselines on PHOENIX14T, GSL, and ISL, and *narrows* (but does not close) the gap on LSA-T, consistent with trends observed as SLT scales (Zhang et al., 2024).

[1]Url anonimized for review purposes.

2 Related Work

Research on Sign Language Translation (SLT) began with the introduction of RWTH-PHOENIX-Weather 2014T and the first end-to-end baselines by Camgoz et al. (2018), which established the now-standard formulation of translating continuous sign video directly into spoken/written text. Subsequent transformer-based architectures advanced the state of the art by better modeling long-range temporal dependencies and jointly learning recognition and translation objectives (Camgoz et al., 2020). More recently, efforts to *scale* SLT in both data and directions highlighted that broader, multi-domain supervision can yield sizeable gains, especially when training setups move beyond a single sign language and a single target (Zhang et al., 2024). Nevertheless, the field has remained predominantly *monolingual*, in large part because sign corpora are scarce, heterogeneous, and difficult to align across languages, which complicates the construction of unified training pipelines and fair evaluation.

In contrast, multilingual modeling has been a defining trend in spoken/written neural machine translation (NMT). A single Transformer with a shared subword vocabulary and simple target-language control tokens can successfully learn many-to-many mappings, facilitate transfer for low-resource pairs, and even enable zero-shot generalization (Johnson et al., 2017). This paradigm naturally motivates multilingual SLT, where the model could amortize learning across sign languages that share articulatory patterns (e.g., hand trajectories, mouthings) or pragmatic structures, while still specializing to language-specific phenomena through conditioning.

Early steps toward multilingual SLT made this connection explicit. Yin et al. (2022) proposed and systematically explored many-to-one, one-to-many, and many-to-many setups, reporting that naive full sharing can cause interference, and that architectural controls (e.g., language-aware routing) help balance sharing versus specialization. Building on this line, Zhang et al. (2025) showed that automatically clustering sign languages into families and training family-specific models can further mitigate negative transfer while preserving the benefits of multilingual supervision. In parallel, work on

scaling SLT emphasized the importance of enlarging both data and translation directions, reinforcing that multilinguality, when properly managed, acts as both regularizer and data multiplier (Zhang et al., 2024). Against this backdrop, our study intentionally opts for a simpler design choice: a fully shared, standard Transformer without routing or family modules, paired with target-language tokens, to isolate how far basic parameter sharing can go in a *multi-target* configuration where each sign language maps to its native spoken language (akin to multilingual NMT) (Johnson et al., 2017).

Finally, the feasibility of multilingual SLT also hinges on the availability of diverse corpora beyond PHOENIX14T. Recent datasets such as the Greek Sign Language Dataset (Adaloglou et al., 2020), LSA-T for Argentinian Sign Language (Bianco et al., 2023), and ISLTranslate for Indian Sign Language (Joshi et al., 2023) broaden the linguistic and domain coverage for SLT research. Yet these resources differ in annotation conventions, domains, and difficulty, complicating joint training. This motivates standardized preprocessing layers and unified data schemas, which we leverage to train and evaluate a single pose-based model across multiple sign languages within one coherent framework.

3 Methodology

3.1 Datasets and Data Processing

Our study spans four SLT corpora with diverse languages, domains, and collection protocols: RWTH-PHOENIX-Weather 2014T (Camgoz et al., 2018), Greek Sign Language Dataset (GSL) (Adaloglou et al., 2020), LSA-T (Bianco et al., 2023), and ISLTranslate (Joshi et al., 2023). To make joint training feasible and comparable across languages, we standardize all datasets through a unified schema that normalizes splits, text preprocessing, and video-to-sequence conversion.

Concretely, videos are sampled at a consistent frame rate and processed with MediaPipe to extract 2D landmarks for hands, upper body, and selected facial regions (Lugaresi et al., 2019). We apply temporal smoothing and torso-based normalization to reduce jitter and scale variance, then select a subset of $\sim$150 features per frame (prioritizing hands/arms and a small set of facial cues) that best capture manual articulations and grammatical markers. Text targets are normalized and tokenized with a shared subword vocabulary. Figure 1 illustrates how the multilingual training set is formed by concatenating all corpora and converting each video to a pose-keypoint sequence, and, as a side benefit, using pose keypoints instead of raw frames also reduces sensitivity to dataset-specific nuisances (e.g., backgrounds, lighting, clothing, camera/viewpoint, signer appearance), promoting more invariant cross-corpus sharing while preserving motion/configuration cues.

3.2 Training Procedure

We adopt a two-stage schedule designed to leverage cross-lingual transfer while preserving language-specific nuances:

Stage 1 (Multilingual pre-training): we train a single fully shared model on the union of all datasets. To avoid overfitting to high-resource subsets, mini-batches are balanced by oversampling lower-resource languages, and early stopping is triggered on a macro-averaged validation BLEU across languages. The objective is standard cross-entropy over subword targets; we do not use gloss supervision.

Stage 2 (Language-specific fine-tuning): starting from the multilingual checkpoint, we fine-tune one model per language with a lower learning rate, which reliably recovers (and sometimes surpasses) the monolingual baselines. Throughout, the target-language token conditions the decoder so that the same parameters handle DGS$\rightarrow$German, GSL$\rightarrow$Greek, LSA$\rightarrow$Spanish, and ISL$\rightarrow$English within one architecture (Johnson et al., 2017). The full workflow is summarized in Figure 2.

3.3 Model Architecture

Our model builds on **Signformer** (Yang, 2024), a compact Transformer sequence-to-sequence architecture. We replace the original frame-based convolutional tokenization with a pose-based encoder: each frame's selected keypoints (hands, upper body, facial cues) are concatenated into a vector of dimension $d_{in} \approx 150$, normalized, and linearly projected to the model embedding space. Unlike multilingual SLT systems that introduce language-specific routing or adapters (Yin et al., 2022), we keep all parameters shared, emphasizing simplicity and parameter efficiency. Figure 3 illustrates the model's architecture.

Beyond efficiency, the pose-based encoder acts as an *inductive bias* toward signer and background invariant features, encouraging cross-lingual shar-

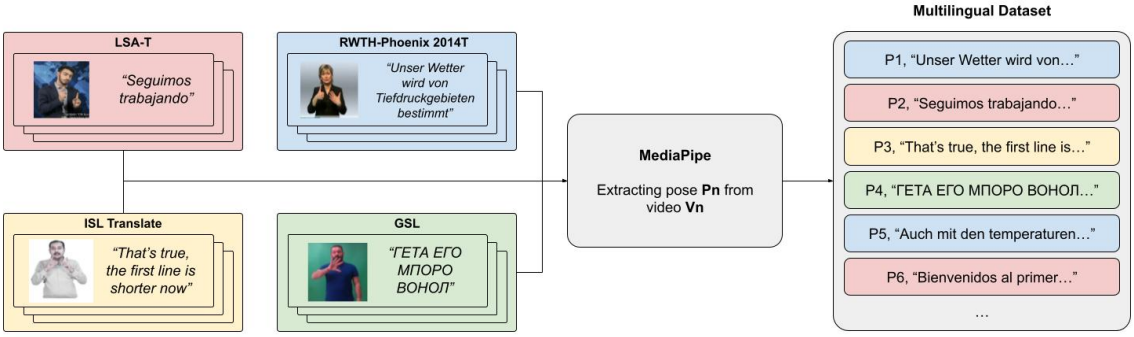

Figure 1: Multilingual dataset construction. Each corpus (PHOENIX14T, GSL, LSA-T, ISLTranslate) is standardized via a unified schema, then each video is converted into a sequence of MediaPipe keypoints (hands/body/face). The resulting pose sequences are concatenated into one multilingual training set with target-language tokens for multi-target decoding.

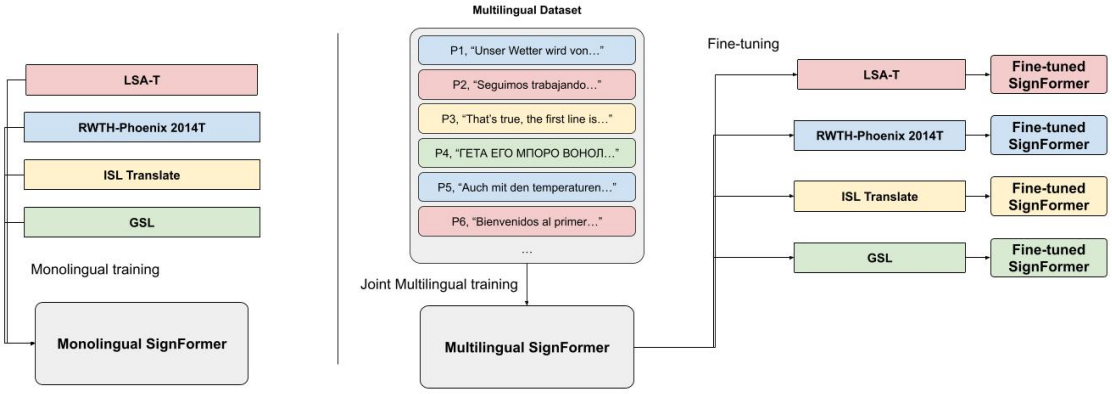

Figure 2: Monolingual training vs Two-stage multilingual training. *Stage 1:* joint pre-training of a fully shared Signformer on the concatenation of PHOENIX14T, GSL, LSA-T, and ISLTranslate with target-language tokens. *Stage 2:* light fine-tuning on each language's data starting from the multilingual checkpoint.

ing without overfitting to visual artefacts that differ across datasets.

4 Experiments and Results

We evaluate three training regimes: (i) **Monolingual baselines**—one pose-based *Signformer* per dataset; (ii) a **Multilingual joint** model trained naively on the concatenation of all corpora; and (iii) **Multilingual + fine-tuning**, where the joint model is lightly adapted to each language. We report case-insensitive BLEU-4 (Papineni et al., 2002), following standard SLT practice (Camgoz et al., 2018, 2020). Table 1 summarizes results for all four datasets.

Two clear trends emerge. First, *naive* joint training under a fully shared parameterization incurs sizeable drops relative to monolingual training (PHOENIX14T: -5.29; GSL: -31.31; LSA-T: -0.70; ISL: -2.02 BLEU), indicating capacity dilution and cross-language interference when mixing heterogeneous sign languages without stronger sharing controls. Second, the *two-stage* schedule is crucial: brief, low–learning-rate fine-tuning

Dataset	Monolingual	Joint	+Fine-tune
PHOENIX14T (DGS→De)	9.56	4.27	**9.71**
GSL Dataset (GSL→Gr)	94.38	63.07	**95.12**
LSA-T (LSA→Es)	**1.18**	0.48	0.94
ISL-Translate (ISL→En)	2.61	0.59	**2.71**

Table 1: BLEU-4 on test sets for monolingual baselines, a single multilingual joint model, and multilingual pre-training followed by language-specific fine-tuning. Best per row in **bold**.

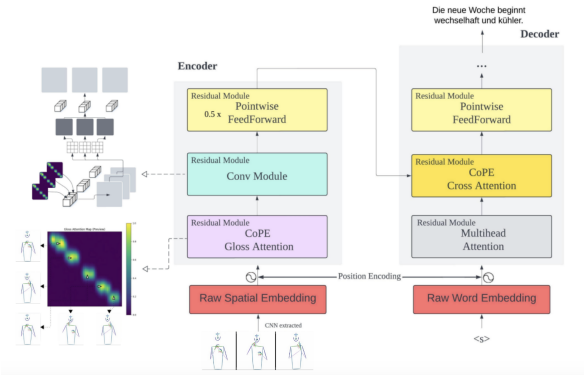

Figure 3: Overview of the adapted *Signformer* architecture (originally taken from (Yang, 2024)) for multilingual SLT using pose keypoints as input. Instead of frame-based visual tokens, each frame's concatenated hand, upper-body, and selected facial landmarks (after normalization and linear projection) feed the encoder. A shared decoder, conditioned on a target-language token, generates the translation.

largely restores and, on *three* datasets, *surpasses* monolingual performance (PHOENIX14T $+0.15$, GSL $+0.74$, ISL $+0.10$ vs. monolingual), while LSA-T remains challenging (joint $\rightarrow$ FT: $+0.46$, ending -0.24 below monolingual). These outcomes mirror multilingual MT and SLT scaling results—multilingual pre-training acts as a regularizer and data multiplier, but sensitive adaptation is required to realize gains across languages and domains (Johnson et al., 2017; Zhang et al., 2024).

Dataset complexity and representation effects. The spread in BLEU-4 reflects intrinsic differences across corpora. GSL's curriculum-oriented content and constrained phrasing may partly explain its very high scores, whereas LSA-T's news-style, open-domain content, signer variability, and potential annotation/pose-estimation noise make it considerably harder. Moreover, pose-based inputs—while enabling compact, deployable models—trade some fine-grained appearance cues (e.g., subtle handshapes, facial expression nuances) for efficiency, which can widen the gap to video-based SOTA on the most challenging settings (Yang, 2024). Still, the fact that PHOENIX14T and GSL not only recover but slightly surpass monolingual baselines after multilingual pre-training suggests that shared motion/configuration patterns are learnable with keypoints when paired with light language-specific adaptation.

5 Conclusion

We presented a multi-target multilingual SLT system that translates DGS$\rightarrow$German, GSL$\rightarrow$Greek, LSA$\rightarrow$Spanish, and ISL$\rightarrow$English within a single, fully shared Transformer, enabled by a unified data layer and pose-based inputs. Naive joint training alone is insufficient—performance drops on all four datasets—but a simple two-stage schedule (multilingual pre-training followed by brief language-specific fine-tuning) reliably recovers and *surpasses* monolingual baselines on PHOENIX14T, GSL, and ISL, while narrowing (though not closing) the gap on LSA-T. These findings echo multilingual MT and recent SLT scaling results: cross-lingual transfer is beneficial, but careful adaptation is necessary to mitigate interference (Johnson et al., 2017; Zhang et al., 2024).

Relative to prior multilingual SLT that commonly evaluates many-to-one into a single target language, our study emphasizes a *multi-target* configuration aligned with each dataset's native spoken language and demonstrates that a compact, pose-based *Signformer* can serve as an effective backbone for this setting. While pose inputs may underperform on unconstrained domains like LSA-T, they enable lightweight, privacy-friendly models.

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*.

Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. 2020. Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*.

Pedro Dal Bianco. 2025. Slt datasets downloader. GitHub repository. Accessed 2025-09-23.

Pedro Dal Bianco, Gastón Ríos, Franco Ronchetti, Facundo Quiroga, Oscar Stanchi, Waldo Hasperué, and Alejandro Rosete. 2023. 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*, pages 293–304. Springer, Cham.

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 (CVPR)*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Abhinav Joshi, Susmit Agrawal, and Ashutosh Modi. 2023. Isltranslate: Dataset for translating indian sign language. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10466–10475.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Amit Moryossef. 2024. Optimizing hand region detection in mediapipe holistic full-body pose estimation to improve accuracy and avoid downstream errors. *arXiv preprint arXiv:2405.03545*.

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.

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

Aoxiong Yin, Zhou Zhao, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. 2022. Mlslt: Towards multilingual sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Biao Zhang, Garrett Tanzer, and Orhan Firat. 2024. Scaling sign language translation. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*.

Ruiquan Zhang, Cong Hu, Pei Yu, and Yidong Chen. 2025. Improving multilingual sign language translation with automatically clustered language family information. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*.