A Transformer-Based Framework for Greek Sign Language Production using Extended Skeletal Motion Representations

Chrysa Pratikaki
Robotics Institute, Athena Research Center
Athens, Greece
National Technical University of Athens
Athens, Greece

Panagiotis Paraskevas Filntisis
Robotics Institute, Athena Research Center
Athens, Greece

Athanasios Katsamanis
Athena Research Center
Athens, Greece

Anastasios Roussos
Institute of Computer Science, Foundation for Research & Technology - Hellas (FORTH)
Heraklion, Greece

Petros Maragos
Robotics Institute, Athena Research Center
Athens, Greece
National Technical University of Athens
Athens, Greece

Abstract

Sign Languages are the primary form of communication for Deaf communities across the world. To break the communication barriers between the Deaf and Hard-of-Hearing and the hearing communities, it is imperative to build systems capable of translating the spoken language into sign language and vice versa. Building on insights from previous research, we propose a deep learning model for Sign Language Production (SLP), which to our knowledge is the first attempt on Greek SLP. We tackle this task by utilizing a transformer-based architecture that enables the translation from text input to human pose keypoints, and the opposite. We evaluate the effectiveness of the proposed pipeline on the Greek SL dataset Elementary23, through a series of comparative analyses and ablation studies. Our pipeline's components, which include data-driven gloss generation, training through video to text translation and a scheduling algorithm for teacher forcing - auto-regressive decoding seem to actively enhance the quality of produced SL videos.

CCS Concepts

• **Human-centered computing** → **Assistive systems and tools**; • **Computing methodologies** → **Transformer networks**; **Gesture recognition**.

Keywords

Sign Language Production, Transformer Networks, Gesture Recognition

PETRA '25, June 25–27, 2025, Corfu Island, Greece

ACM ISBN 979-8-4007-1402-3/25/06
https://doi.org/10.1145/3733155.3733204

ACM Reference Format:
Chrysa Pratikaki, Panagiotis Paraskevas Filntisis, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. 2025. A Transformer-Based Framework for Greek Sign Language Production using Extended Skeletal Motion Representations. In *The PErvasive Technologies Related to Assistive Environments (PETRA '25), June 25–27, 2025, Corfu Island, Greece.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3733155.3733204

1 Introduction

To address communication barriers between the DHH (Deaf and Hard-of-Hearing) and the hearing communities, the field of Sign Language Processing has emerged at the intersection of linguistics, computer vision, and machine learning. Sign Language Processing encompasses a variety of tasks, such as automatic translation and production of sign language, with the most critical components of an effective sign language system being Sign Language Translation (SLT), and Sign Language Production (SLP). In this paper, we primarily focus on Sign Language Production (SLP), which involves generating accurate sign language sequences from a given text input. Specifically, we approach the production of sign language videos by proposing a transformer-based method to generate extended skeletal representations from text.

SLP systems have primarily relied on basic animation techniques or rule-based models, which often fail to capture the subtleties of human motion and natural language. Recent advancements in deep learning, particularly neural machine translation models and generative networks, have opened new possibilities for generating more photorealistic sign language content. Despite these advances, current solutions are still in their early stages, with significant room for improvement in the fluidity and accuracy of the produced sign language videos. In this work, we utilize transformer architectures aiming to address the existing SLP limitations. Our contributions can be summarized as follows:

(1) We present the first SLP system for the Greek Language based on deep learning.

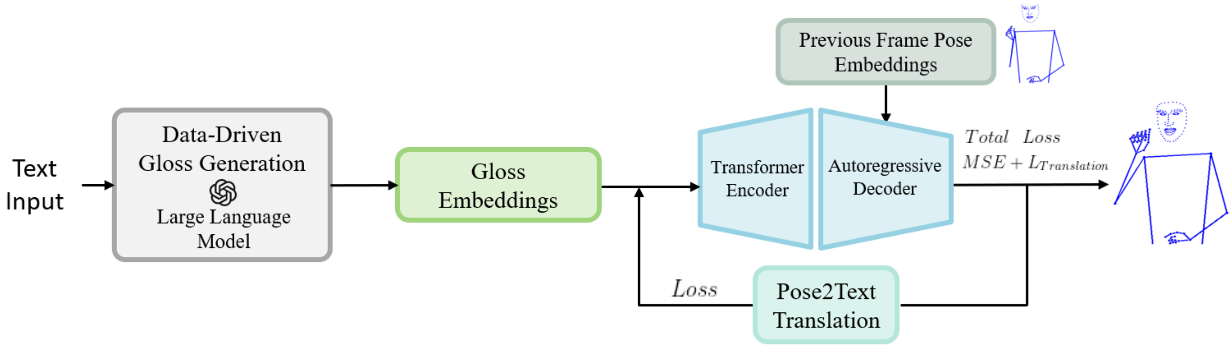

Figure 1: Overview of the proposed architecture: Given a text sentence as input, our SLP pipeline generates the corresponding sign language sequence. During training, the Encoder-Decoder structure learns through a sum of MSE Regression Loss (between frames) and CTC (pose-to-text) Loss. Optionally, training can happen using data-driven generated glosses to limit lexical diversity.

(2) Our method incorporates components such as the use of a Pose-to-Text loss during training and SL gloss generation through text transcriptions, which help the quality of the generated sign poses. We also propose a scheduling algorithm that alternates between using teacher forcing and auto-regressive decoding during training.

(3) We conduct experiments on the publicly available Greek Sign Language dataset, Elementary23 [20]. Through extensive quantitative evaluations and ablation studies we highlight the strengths and weaknesses of the proposed transformer-based architecture, achieving significant improvements on the quality of the SLP results.

2 Related Work

Sign Language Recognition and Translation: Sign Language Recognition (SLR) focuses on extracting meaningful features from sign language videos and classifying them into discrete sign representations (glosses), while Sign Language Translation (SLT) is defined as the translation of sign language videos directly into spoken language. Early work on Sign Language Processing ([17], [11]) addresses SLR as a computer vision problem, focusing on enhancing hand recognition accuracy, by utilizing statistical subunits and lexicons. Many recent works have tackled both SLR and SLT tasks using a variety of deep learning approaches, including RNNs ([1], [2]), LSTMs ([6]), GRUs ([8]), and Transformers ([4], [3]), after using CNNs for spatial feature extraction or Pose Estimation networks. Camgöz et al. [2] formalized SLT as a sequence-to-sequence (seq2seq) learning problem. This approach employs CNNs for spatial feature extraction from sign language videos, which are then fed into an attention-based encoder-decoder framework to generate spoken language translations. These experiments were made on three different pipelines, gloss-to-text, sign-to-text and sign-to-gloss-to-text, which uses gloss annotations as an intermediate layer. In another work, Camgöz et al. [4] used transformer models for both the recognition and translation pipelines. The encoders process sign video sequences to produce embeddings that capture both spatial and temporal features, while the decoders generate spoken language sentences. CTC Loss is used to facilitate learning without explicit alignment data, tying the recognition of sign glosses to the generation of text. Experimental results of the previously mentioned works prove that using gloss information as an intermediate step to spoken language translation improves the performance of the model, however relying on gloss annotations can be limiting on larger datasets since they require professional annotation.

Sign Language Production: While Sign Language Translation has seen considerable progress, Sign Language Production remains under-explored, with a need for significant breakthroughs. Early works on SLP primarily relied on phrase lookup, direct sentence matching, and computer-generated avatar sign videos to produce realistic animated outputs, such as Tessa (BSL) [14] and Simon (Sign Supported English) [10]. Recent advancements [12], [15], [13], [16], have redefined SLP as a **Neural Machine Translation** (NMT) task, leveraging sequence-to-sequence models to generate 2D pose sequences from text embeddings. Saunders et al. [12] pioneered a Transformer-based architecture for end-to-end SLP, employing a dual-transformer approach. Their Symbolic Transformer encodes text, while the Progressive Transformer generates continuous frame sequences, marking a significant step forward in automating and enhancing SLP. Several other works [13], [19], [18], specialize on the photorealistic aspect of SLP and aim to synthesize and also anonymize realistic SL videos.

Datasets: Most mentioned works on SLT and SLP conduct their experiments on the publicly available **PHOENIX14T** dataset [9]. This dataset includes a total of 8257 sequences performed by 9 signers along with their gloss annotations. Its relatively limited vocabulary of 1066 sign glosses allows for high quality SLT and SLP results. On the other hand the **Elemntary23** dataset [20], is a recent GSL dataset, which contains annotations of the first three classes of Greek Elementary school books in all subjects. The Greek Language subset containts 9499 videos with a vocabulary of 14345 words, while the Maths subset containts 6583 videos with a vocabulary of 6457 words.

Pose Estimation: Recent advances in the field of computer vision and pose estimation, make it possible to generate 2D or 3D landmarks from RGB images. Open Pose ([5],[21]) is one of the first and most popular frameworks for human pose estimation and is mostly used in the previously mentioned works on SLT and SLP. MediaPipe (MP) is an open-source framework for constructing multi-modal and cross-platform machine learning pipelines, supporting a broad range of applications, including pose and face detection. A particular implementation of MediaPipe is the MediaPipe Holistic model [7], which we particularly use in this work. MP Holistic employs a graph-based pipeline that processes different regions of interest (ROIs) within an image to estimate a total of up to 543 landmarks, which include 33 body pose landmarks, up to 468 facial landmarks, and 42 hand landmarks (21 per hand).

3 Methodology

Given a **text** sentence as **input**, our SLP pipeline generates the corresponding 2D sign language sequence in the form of extended skeleton representations, which include hand, face and body landmarks. An overview of the proposed architecture is presented in Fig. 1. It consists of four main components: Feature Extraction, Gloss Extraction, Auto-regressive Decoding, and Pose-to-Text Translation, which are analyzed in the following sections.

3.1 Feature Extraction

We first extract the skeleton pose sequences from Elementary23 SL videos using MediaPipe (MP) Holistic [7]. In order to expedite the training process, we sub-sample both the pose and face mesh landmarks. For the pose keypoints, we select the 8 points shown in Figure 2 which include the body parts necessary for a SL video, such as the torso, elbows and wrists. For the face landmarks, we choose 141 instead of 468 face keypoints, which contain all the necessary face information, such as the mouth, eyes, nose and face perimeter. For each hand, we keep all 21 landmarks. This brings us to a total of 191 landmarks, instead of the original MP 543 landmarks, which is a substantial reduction to nearly one third. Finally, the total extracted landmark sequence for each frame is defined as follows:

$$\mathbf{P}_f = [\mathbf{a}_{left\ hand} || \mathbf{a}_{right\ hand} || \mathbf{a}_{face} || \mathbf{a}_{pose} || c_f] \qquad (1)$$

where P_f is the landmark sequence for the f-th frame, c_f is the counter value ranging from 0 to 1 that indicates the relevant frame posisition and || the concatenation symbol.

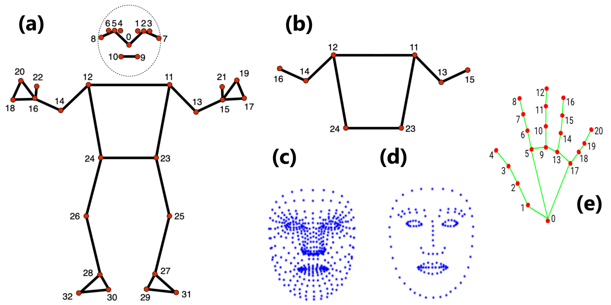

Figure 2: Extended Skeleton Representation based on MediaPipe Holistic [7]: (a) Original 33 MP pose landmarks. (b) Selected 8 MP pose landmarks for SLP. (c) Original 478 MP face landmarks. (d) Selected 141 MP face landmarks for SLP. (e) MP hands.

3.2 Text to Video Transformer Module

After extracting the pose sequences from the dataset, we employ a transformer-based architecture to tackle the Sign Language Production (SLP) task, and ultimately transform text sentences to sign sequences. The key distinction between our approach and a classic encoder-decoder architecture is that the decoding process happens auto-regressively, meaning the model produces a sign pose frame at each time-step given the text embeddings and the previously generated pose embeddings. The training objective of the transformer module is concluded by regressively calculating the MSE between the ground-truth $y^*_{1:F}$ landmark sequence and the predicted landmark sequence $\hat{y}_{1:F}$, with F being the total number of frames. Fig. 3 shows the **text2pose** Transformer architecture.

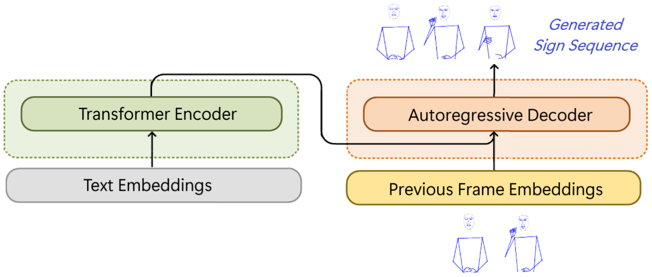

Figure 3: Proposed Sign Language Production Transformer

3.3 Teacher Forcing vs Auto-regressive Decoding

In previous methods [12], transformer models were trained using teacher forcing. This approach involves providing the model with the ground truth spatial embeddings from the previous frame during sequence generation. By using the correct embeddings as input, this method enables parallel training with known outputs. While teacher forcing has demonstrated satisfactory results on limited vocabulary datasets such as PHOENIX14T [4, 12], it struggles with the broader and more diverse Greek Sign Language dataset. In general, while teacher forcing provides better training stability and ensures alignment between inputs and outputs—particularly in the earlier stages of training—it suffers from error compounding during inference, as the network is unable to recover from its own prediction errors.

On the contrary, autoregressive training generates frame sequences sequentially during training as well. In this approach, the model predicts each frame by conditioning on the spatial embeddings it has previously generated. Before applying the MSE loss, the entire sign sequence is generated from the text embeddings, effectively mimicking the inference process. This allows the model to learn to correct its own errors rather than relying on ground-truth inputs. However, this training process is considerably more time-consuming than teacher forcing due to the sequential nature of frame generation.

To balance efficiency and effectiveness, we employed a hybrid approach, training the model using teacher forcing and autoregressive generation for half of the epochs each. Specifically, we began training with teacher forcing to leverage its stability and strong input-output alignment during the critical early stages of training. This ensures the model effectively learns the foundational relationships in the data. We then switched to autoregressive training, allowing the model to learn to correct its own errors and better handle the challenges of inference. This strategy combines the strengths of both methods, resulting in improved performance compared to using either method independently. Quantitative results comparing these approaches are presented in Section 4.

3.4 Video to Text Translation Module

An integral component of our training pipeline is the implementation of the **pose-to-text** loss. This approach entails the pre-training of a distinct Translation model that maps the 2D sign sequences to text, which is subsequently utilized during the training of the text2pose (forward process) model. The objective is to enhance accuracy and prevent the model from regressing to mean pose, which often happens when only training with MSE loss, and also prove its ability to reinforce the quality of the forward translation. The translation loss, essential for both training and evaluation, is formulated following [4] as follows:

$$L_T = 1 - \prod_{u=1}^{U} \sum_{d=1}^{D} p(\hat{w}_u^d) p(w_u^d | h_u) \quad (2)$$

where $p(\hat{w}_u^d)$ is the probability of word w^d at decoding step u, while D is the vocabulary size. $\prod_{u=1}^{U} p(w_u^d | h_u)$ is calculated by sequentially applying CTC Loss on a frame level for each word.

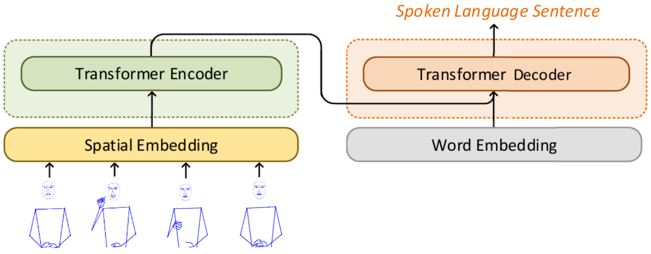

Figure 4: Proposed Sign Language Translation Transformer

In order to implement the pose-to-text translation model we built on the state-of-the-art, publicly availiable network Sign Language Transformers by Camgoz et al. [4]. Simplifying the overall training process that performs both SL recognition and translation, we keep solely the translation loss objective, aiming to achieve the desired results through a direct sign2text model. The complete architecture of the **pose-to-text** module is shown in Figure 4.

3.5 Gloss Extraction as an intermediate step

Next, we explored the use of off-the-shelf large language models (LLMs) to automatically generate gloss annotations for the SL dataset. This approach effectively reduces the lexical diversity of the dataset by condensing commonly used words, such as articles and connective phrases, while preserving the overall meaning of the sentences. Given the established benefits in previous literature, where gloss annotations as an intermediate step have been shown to enhance model performance [12], [15], we anticipate observing similar improvements in our experiments. Shown below in table 1 are Gloss Generation examples from the Elementary23 Dataset using the GPT-4o API.

Prompt	Transform this Greek sentence into Greek Sign Language gloss: "I complete the table by first estimating the values approximately and then checking my calculations"
Gloss	COMPLETE TABLE ESTIMATE FIRST VALUES APPROXIMATELY CHECK THEN CALCULATIONS
Prompt	Transform this Greek sentence into Greek Sign Language gloss: "The line of symmetry divides shhapes into two equal parts"
Gloss	LINE SYMMETRY DIVIDES SHAPE TWO EQUAL PARTS
Prompt	Transform this Greek sentence into Greek Sign Language gloss: "I observe and continue the patterns"
Gloss	OBSERVE CONTINUE PATTERNS

Table 1: Examples of the text-to-gloss sequence translation that we adopt, based on LLMs. Note that the original sentences and gloss outputs are in Greek, however we present here the English translations.

4 Experiments and Evaluation

In this section, we provide extensive comparative analyses and ablation studies on the Elementary23 dataset. First, we explore how the model performs on Signer-dependent subgroup of the Maths subset (for the two most frequently appearing Signers in Elementary 23 Maths subset). Next we explore how the model performs on entire sections of the dataset, with a class-related theme (i.e. Maths Subset, Greek Language Subset). Following [12], [13], we perform evaluation using the NLP metrics BLUE-4 and ROUGE-L, and also DTW (Dynamic Time Wrapping) for measuring similarity between the produced and ground-truth sign sequence.

4.1 Dataset

As mentioned in section 2, the Elementary23 dataset [20] contains annotations of the first three classes of Greek Elementary school books in all subjects, with a large vocabulary exceeding 30,000 words. In our work, we focus on The Greek Language subset which contains 9499 videos with a vocabulary of 14345 words, and the Math subset which contains 6583 videos with a vocabulary of 6457 words. Specifically for the Math subset, we begin our evaluation in subsection 4.3, by training the SLP pipeline on the two most prominent signers, referred to as Signer A and Signer B, who appear in 3,476 and 746 videos, respectively. Then, in subsection 4.4, we generalize our training process across all signers to achieve a more holistic result. Table 3 visualizes the size and vocabulary of each subset used.

4.2 Evaluation of the sign-to-text module

First, we evaluate the Sign Language Translation (sign-to-text) Module. The sign-to-text module is crucial for our pipeline, as it's used during SLP evaluation in the following sections, as well as during text-to-sign SLP training. Table 2 shows that we achieve a BLUE-4 score of 7.69 on the Math and 5.52 on the Greek subset, which is quite promising and close to the 6.67 mentioned in the original paper [20] on the SLT task.

	Dev		Test	
Sign-to-Text Method	**BLEU-4↑**	**ROUGE↑**	**BLEU-4↑**	**ROUGE↑**
Voskou et al. [20], trained on entire Elementary23	6.67	-	5.69	-
Ours, trained on Elementary23 Math	7.58	15.11	7.69	15.26
Ours, trained on Elementary23 Greek	5.63	14.56	5.52	14.23

Table 2: Evaluation of the sign-to-text module on the Elementary23 SL Dataset. Please note that results are not directly comparable due to differences in the test sets.

		Videos	# Words
Math	**Signer A**	3473	3654
	Signer B	746	1059
Greek	**Signer A**	2467	4749
	Signer B	1927	3535

Table 3: Elementary23 subsets

4.3 Evaluation of signer-specific training

To begin our SLP evaluation, we performed experiments on the Mathematics subset of the Elementary23 dataset, on the two most frequently appearing signers, Signer A and Signer B. To assess the model's ability to generalize across different signers, we performed two separate training sessions: one focused exclusively on Signer A and the other on Signer B. For evaluation, we alternated between their respective test sets to measure cross-signer performance. Results are shown in table 4. We clearly see that the model fails to produce accurate signs when the "wrong" test set is used. These findings suggest that the model struggles to generalize across signers, likely due to differences in signing styles or vocabulary correlations unique to individual signers. To address this limitation, we proceed to train our models on larger sections of the Elementary23 dataset, emphasizing the need for more generalized training approaches.

	Test - Signer A		**Test - Signer B**	
	BLEU-1↑	**BLEU-4↑**	BLEU-1↑	**BLEU-4↑**
Train - Signer A	17.05	**5.02**	5.93	0.00
Train - Signer B	6.29	1.18	21.87	**6.69**

Table 4: Ablation Study on the Elementary23 Greek Language SL Dataset. Best-performing results are highlighted in bold, while failure scores in the case of swapped signer test scores are shown in red.

4.4 Signer Independent Studies

Next, we focus on conducting experiments on entire sections of the Elementary23 dataset, disregarding the fact that videos are filmed with different signers. We specifically choose the entire Math subset and the Greek Language subset. Our first ablation study, shown in table 6, compares training with Teacher Forcing (TF), Auto-regressive Decoding (AD), and their combination (TF+AD), underlining the benefits of employing a hybrid approach. While Auto-regressive Decoding achieves significantly higher BLEU-4 and ROUGE scores (5.4 and 14.5 on the dev set, respectively) compared to Teacher Forcing (1.69 and 8.52), the hybrid TF+AD model provides balance between computational efficiency and predictive accuracy. Notably, the hybrid model achieves the highest overall performance, both in the Greek and Math subsets, validating the importance of alternating decoding strategies during training.

		Dev	**Test**
$pose - to - text\ Loss$	$Gloss$	**BLEU-4↑**	**BLEU-4↑**
✗	✗	**4.17**	4.15
✗	✓	3.56	3.44
✓	✗	**4.42**	**4.55**
✓	✓	4.06	**4.32**

Table 5: Ablation Study on the Elementary23 Greek Language Dataset. We highlight best performing scores in both dev and test sets.

			Time/	**Dev**	**Test**
	Method	Epochs	Epoch (s)	**BLEU-4↑**	**BLEU-4↑**
Greek	Teacher Forcing, [12]	2500	5	0.49	0.35
	Autoregressive Dec	2500	30	4.3	4.13
	TF + AD	1250, 1250	5, 30	**4.67**	**4.46**
Math	Teacher Forcing, [12]	2500	5	1.69	1.46
	Autoregressive Dec	2500	30	5.4	5.3
	TF + AD	1250, 1250	5, 30	**5.69**	**5.59**

Table 6: Ablation Study on the Elementary23 Greek (Top) and Math (Bottom) SL Dataset. Best TF+AD results are highlighted in bold. Teacher Forcing (TF) method can be considered equivalent to the Progressive Transformers (PT) work [12].

Our next ablation study, shown in table 5, on the Elementary23 Greek Language Subset, shows that the inclusion of pose-to-text Loss and Gloss annotations possessively affects on performance. Although BLEU-4 scores improve independently (4.42 dev, 4.55 test), the combination with Gloss yields mixed results, slightly reducing BLEU-4 on the dev set (4.06) but improving on the test set (4.32). This interplay suggests that while gloss annotations simplify linguistic diversity, over-reliance on glosses can limit adaptability. Through these ablation studies we structure a useful evaluation process of our architecture's components, demonstrating how they enhance SLP performance.

4.5 Experimental Setup

All models have been trained using 2-layer transformers with 4 attention heads, embedding dimension 512 and all weights are initialized with Xavier initialization. We trained all modesl using the Adam optimizer with a learning rate of 1e-4, a batch size of 32,

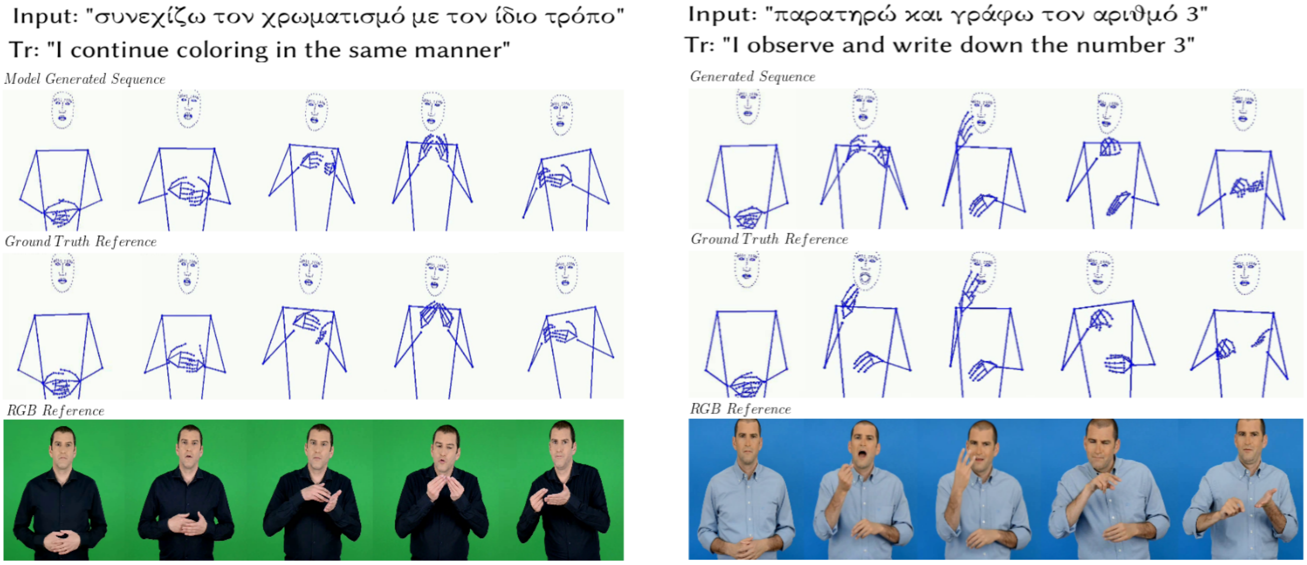
Input: "συνεχίζω τον χρωματισμό με τον ίδιο τρόπο"
Tr: "I continue coloring in the same manner"
Model Generated Sequence
Ground Truth Reference
RGB Reference
Input: "παρατηρώ και γράφω τον αριθμό 3"
Tr: "I observe and write down the number 3"
Generated Sequence
Ground Truth Reference
RGB Reference

Figure 5: Sample (test set) visualizations of our SLP method. Top to bottom: Text inputs, 2D generated sign sequence from text embeddings, ground-truth sequence reference, RGB reference. Figures are best viewed in video form.

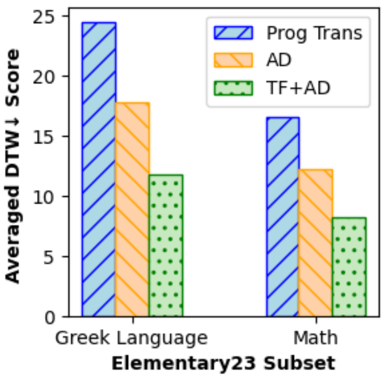

Figure 6: Comparison of the averaged DTW results on the Math and Greek Test Subsets. Again the hybrid combination of teacher forcing and auto-regressive decoding during training significantly improves sequence alignment.

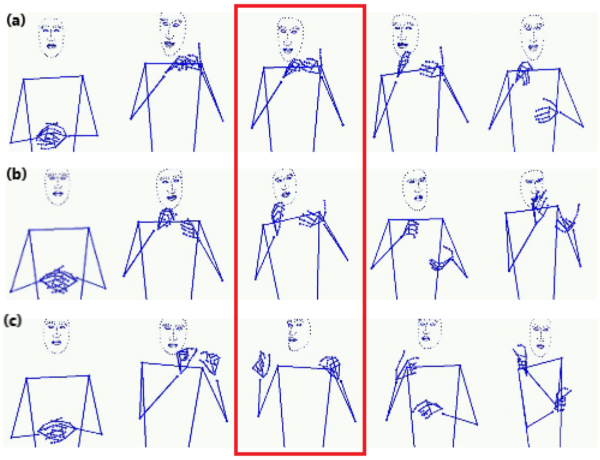

Figure 7: Sample visualization of the effect of the pose-to-text Loss. Top to bottom: (a) 2D Pose w/o pose-to-text Loss, (b) 2D Pose with pose-to-text Loss, (c) ground-truth sequence reference. When used, the generated poses show greater movement variability and regress less on mean pose.

and early stopping based on validation loss. For compatibility across pipelines, all SLP and SLT models within a specific data subset (e.g., Math or Greek) adhere to identical architectural specifications.

4.6 Qualitative Results

Lastly, in Figure 5 we showcase **representative results** from our proposed pipeline, evaluated on test sentences from the Elementary23 dataset using our best-performing models.

5 Conclusions

In this paper we presented the first SLP pipeline applied on Greek Sign Language Datasets, actively improving existing architectures through novel components. We presented our best results, which where achieved through the combination of gloss generation, decoding scheduling and pose-to-text translation training. These SLP methods find useful application mainly in the sign language learning process and education. In the future, we aim to expand our work so that it also incorporates a generative module for photorealistic SL video synthesis, as this is considered a necessary component for a SL user. It's finally important to emphasize that SLP models are not intended to replace sign language interpreters. Instead, they serve as a complementary tool, providing an ethical and practical solution for educational purposes.

Acknowledgments

This project is funded by the European Union under Horizon Europe (grant No. 101136568). A. Roussos acknowledges the support by an NVIDIA Academic Hardware Grant Program, which was beneficial in developing and testing the deep learning model for Sign Language Production introduced in this paper.

Funded by
the European Union

References

[1] Kshitij Bantupalli and Ying Xie. 2018. American Sign Language Recognition using Deep Learning and Computer Vision. In *2018 IEEE International Conference on Big Data (Big Data)*.

[2] 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)*.

[3] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Multi-channel Transformers for Multi-articulatory Sign Language Translation. In *Computer Vision – ECCV 2020 Workshops*, Adrien Bartoli and Andrea Fusiello (Eds.).

[4] Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation.

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*.

[6] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. 2017. SubUNets: End-To-End Hand Shape and Continuous Sign Language Recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[7] Google. [n. d.]. MediaPipe Holistic Solution Documentation. https://github.com/google/mediapipe/blob/master/docs/solutions/holistic.md.

[8] Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. 2019. Neural Sign Language Translation Based on Human Keypoint Estimation. *Applied Sciences* 9, 13 (2019).

[9] Oscar Koller, Jens Forster, and Hermann Ney. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding* 141 (Dec. 2015), 108–125.

[10] F. Pezeshkpour, I. Marshall, R. Elliott, and J.A. Bangham. 1999. Development of a legible deaf-signing virtual human. *International Conference on Multimedia Computing and Systems -Proceedings* 1 (1999), 333 – 338. Cited by: 14.

[11] Anastasios Roussos, Stavros Theodorakis, Vassilis Pitsikalis, and Petros Maragos. 2013. Dynamic Affine-Invariant Shape-Appearance Handshape Features and Classification in Sign Language Videos. *Journal of Machine Learning Research* 14, 51 (2013), 1627–1663. http://jmlr.org/papers/v14/roussos13a.html

[12] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. Progressive Transformers for End-to-End Sign Language Production. (2020). http://arxiv.org/abs/2004.14874

[13] 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.

[14] Stephen Cox, Michael Lincoln, Judy Tryggvason, Melanie Nakisa, Mark Wells, Marcus Tutt, and Sanja Abbott. 2003. The Development and Evaluation of a Speech-to-Sign Translation System to Assist Transactions. *International Journal of Human–Computer Interaction* 16, 2 (2003), 141–161. https://doi.org/10.1207/S15327590IJHC1602_02

[15] Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. 2020. Text2Sign: Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks. *International Journal of Computer Vision* (2020).

[16] Stephanie Stoll, Armin Mustafa, and Jean Yves Guillemaut. 2022. There and Back Again: 3D Sign Language Generation from Text Using Back-Translation. *Proceedings - 2022 International Conference on 3D Vision, 3DV 2022*, 187–196. https://doi.org/10.1109/3DV57658.2022.00031

[17] Stavros Theodorakis, Vassilis Pitsikalis, and Petros Maragos. 2014. Dynamic–static unsupervised sequentiality, statistical subunits and lexicon for sign language recognition. *Image and Vision Computing* 32, 8 (2014), 533–549. https://doi.org/10.1016/j.imavis.2014.04.012

[18] Christina O. Tze, Panagiotis P. Filntisis, Athanasia-Lida Dimou, Anastasios Roussos, and Petros Maragos. 2023. Neural Sign Reenactor: Deep Photorealistic Sign Language Retargeting. arXiv:2209.01470 [cs.CV]

[19] Christina O. Tze, Panagiotis P. Filntisis, Anastasios Roussos, and Petros Maragos. 2022. Cartoonized Anonymization of Sign Language Videos. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. 1–5. https://doi.org/10.1109/IVMSP54334.2022.9816293

[20] 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:2310.04753 [cs.CL]

[21] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional pose machines. In *CVPR*.