From text to sign language pose sequences: a Transformer-based approach for the automatic generation of gestures

M. Harrisson HOUENOU SEZAN[1,*,†], Pélagie HOUNGUE[1,†]

[1]*Institute of Mathematics and Physical Sciences (University of Abomey-Calavi), Dangbo, Benin*

Abstract

Access to information remains a major challenge for deaf and hard-of-hearing individuals due to the structural differences between spoken and sign languages. This paper proposes a Transformer-based approach for the automatic generation of sign language pose sequences directly from text without relying on gloss-based intermediate representations. The proposed architecture combines a DistilBERT encoder for contextual text representation with a Transformer decoder for the generation of three-dimensional articulatory poses. Experiments were conducted using the public How2Sign and WLASL datasets, complemented by a local corpus comprising 500 videos collected from 10 signers. The model was evaluated against recurrent baselines including LSTM and GRU architectures. Results show that the proposed approach achieves an accuracy of 70.14%, outperforming GRU (51.38%) and LSTM (38.29%), while also obtaining a lower mean squared error and improved temporal consistency. These findings demonstrate the potential of attention-based architectures for sign language generation and highlight their relevance for improving digital accessibility. Such technologies could contribute to the development of more inclusive communication systems and accessible digital services within smart and sustainable environments.

Keywords

Sign language, Transformers, Deep Learning, Gesture generation, Machine translation, Articulation points

1. Introduction

Communication is a fundamental pillar of human interaction. However, for deaf and hard-of-hearing people, access to information remains severely limited by the gap between spoken languages and sign languages [1]. This language barrier manifests itself in many contexts, particularly in education, access to services, and the dissemination of digital content, thereby contributing to a form of persistent social exclusion.

Beyond its scientific interest, this issue is also part of a broader dynamic of inclusive digital transformation. In the context of smart cities, equitable access to digital services, information platforms, and communication systems is a major challenge for people with disabilities. Machine translation technologies into sign language can thus help reduce communication barriers and promote more inclusive participation by deaf people in modern digital environments.

Unlike spoken languages, sign languages rely on a visual-gestural modality, simultaneously combining hand movements, facial expressions, and spatial configurations. This characteristic makes modeling them particularly complex in an automated setting. Traditional approaches to machine translation cannot be directly applied to this context without adaptation [2]. Machine translation into sign language is currently a rapidly evolving field of research, at the intersection of natural language processing, computer vision, and deep learning.

Existing work in translation into sign language relies primarily on intermediate representations, such as **glosses**, or on **recognition-followed-by- generation systems** [3]. Although these approaches

CITA 2026—Land, Smart Cities, and Sustainable Development, 25–26 June, Cotonou, Benin

*Corresponding author.

†These authors contributed equally.

harrisson.houenousezan@imsp-uac.org (M. H. HOUENOU SEZAN); pelagie.houngue@imsp-uac.org (P. HOUNGUE)

0009-0001-4012-4388 (M. H. HOUENOU SEZAN); 0000-0001-8138-2079 (P. HOUNGUE)

have led to significant advances, they have major limitations, particularly in terms of their dependence on annotations.

In this work, we propose an alternative approach based on the direct generation of pose sequences from text. The idea is to model gestures as articulatory keypoints, thereby capturing the dynamics of movement without explicitly relying on glosses. To achieve this, we rely on a Transformer-style architecture [4], capable of efficiently modeling long-term dependencies thanks to its attention mechanism.

Unlike approaches relying on heavily annotated intermediate representations, our approach aims to directly learn a correspondence between linguistic content and gestural dynamics. This strategy reduces dependence on glosses while preserving the temporal continuity of the generated movements.

The objective of this article is twofold. First, it aims to design a system capable of generating coherent gesture sequences from a text. Second, it seeks to evaluate the contribution of attention-based architectures compared to traditional sequential models such as GRUs and LSTMs.

The central hypothesis of this work is that direct modeling of articulatory poses, combined with an attention-based architecture, allows for better capture of the spatiotemporal dependencies of sign languages than classical sequential approaches.

The main contributions of this work can be summarized as follows:

- Proposal of a text-to-sign translation chain based directly on three-dimensional articulatory representations, without the use of an intermediate representation in the form of glosses;
- Integration of a pre-trained linguistic encoder based on DistilBERT with a Transformer decoder dedicated to generating sign sequences;
- Introduction of a learning function combining spatial accuracy and temporal consistency to improve the fluidity of the generated movements;
- Experimental validation on the public How2Sign and WLASL corpora as well as on a local corpus constructed as part of this study.

The remainder of this article is organized as follows: Section 2 presents the state of the art, Section 3 describes the proposed approach, Section 4 outlines the experimental protocol, section 5 presents the results, section 6 offers an analysis, and section 7 concludes the work.

2. Related Work

Machine translation into sign language is a field of research at the intersection of natural language processing, computer vision, and deep learning. Existing research primarily aims to establish a correspondence between textual or spoken input and a gestural representation that can be understood by deaf and hard-of-hearing people.

Early approaches relied primarily on rule-based systems and statistical models. These methods sought to link linguistic units from spoken languages to predefined gestural representations. However, they had significant limitations when dealing with the structural complexity of sign languages, which are characterized by simultaneous movements, the use of space, and the prominence of non-manual components.

With the emergence of deep learning, several studies have adopted sequence-to-sequence architectures to automate the learning of correspondences between text and gestures. In this context, Camgöz et al. [3] proposed a neural translation approach based on intermediate glosses. Their method improves the alignment between text sequences and gesture sequences, but remains heavily dependent on the availability of costly linguistic annotations.

Subsequently, Camgöz et al. [5] introduced a Transformer-based architecture for sign language recognition and translation. The results show that attention mechanisms enable better modeling of long-term dependencies compared to traditional recurrent architectures.

In the same vein, Li et al. [6] developed the WLASL dataset, which has become an important benchmark for deep learning applied to large-scale gesture recognition. This dataset highlights the importance of having diverse datasets to improve the robustness of translation and recognition models.

Duarte et al. [7] subsequently proposed the How2Sign multimodal corpus, which integrates text, video, and synchronized sign language information for American Sign Language. This dataset facilitates the development of models capable of simultaneously leveraging multiple information modalities in sign language generation and translation tasks.

In parallel with this work, several recent studies have explored the use of representations based on articulatory poses [8]. These approaches use keypoint estimation techniques to represent body and hand movements as spatial coordinates. They offer the advantage of providing a compact representation of gestures while reducing reliance on intermediate linguistic annotations.

The emergence of Transformer architectures [4] has also profoundly influenced recent work in gesture generation. Thanks to the multi-head attention mechanism, these architectures can capture global dependencies between elements in a sequence and efficiently handle complex temporal structures. Their success in fields such as machine translation and speech recognition has driven their adoption in gesture generation tasks.

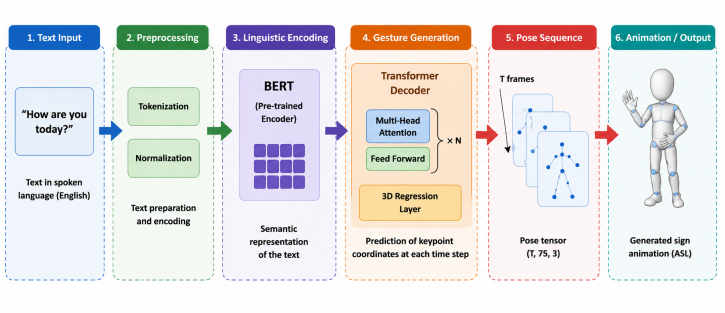

Figure 1: Overview of the proposed text-to-sign pose generation pipeline. Input sentences are first encoded using DistilBERT to obtain contextual representations, which are then processed by a Transformer decoder to generate sequences of three-dimensional articulatory poses without relying on intermediate gloss annotations.

Figure 1 illustrates the general pipeline of the proposed system. Given a textual input, a linguistic encoder extracts a semantic representation, which is then processed by a Transformer decoder to produce a sequence of articulatory poses. This approach allows for the direct generation of gestural trajectories without explicitly relying on intermediate representations such as glosses.

2.1. Discussion on the state of the art

Despite the advances reported in the literature, several limitations remain. Approaches based on glosses require labor-intensive manual annotation and are difficult to scale up. Furthermore, the recurrent architectures used in some studies struggle to effectively capture long-term temporal dependencies and simultaneous interactions between different gestural components.

Furthermore, the lack of large annotated corpora and the high degree of variability among signers remain major challenges for machine translation systems into sign language [1]. Non-manual components, such as facial expressions and head movements, are also insufficiently accounted for in many existing approaches.

These various limitations have prompted research into methods capable of directly modeling articulatory sequences while leveraging the advantages of attention-based architectures. This work falls within this framework, proposing a Transformer-based approach to direct pose generation in order to improve the temporal consistency of the generated gestures and reduce reliance on intermediate annotations.

Unlike several recent approaches that rely on intermediate representations in the form of glosses or multi-stage pipelines, the proposed method adopts a strategy of directly generating articulatory

sequences from text. This choice aims to reduce dependence on computationally intensive linguistic annotations while simplifying the processing chain.

Furthermore, the combined use of a pre-trained language encoder and a specialized Transformer decoder for predicting three-dimensional poses allows for the simultaneous exploitation of both the semantic information in the text and the temporal constraints associated with gestural movements.

3. Proposed Approach

This section presents the proposed model for machine translation of text into sequences of poses corresponding to sign language. The approach adopted is based on a direct modeling of the problem, avoiding the systematic use of intermediate representations such as glosses. The system thus aims to learn a correspondence between an input text sequence and a sequence of articulatory configurations describing the gestures.

The overall architecture follows an **encoder-decoder** framework, in which a semantic representation of the text is used to guide the generation of gesture sequences. To effectively capture temporal and contextual dependencies, we rely on a Transformer-type decoder [9], illustrated in Figure 2.

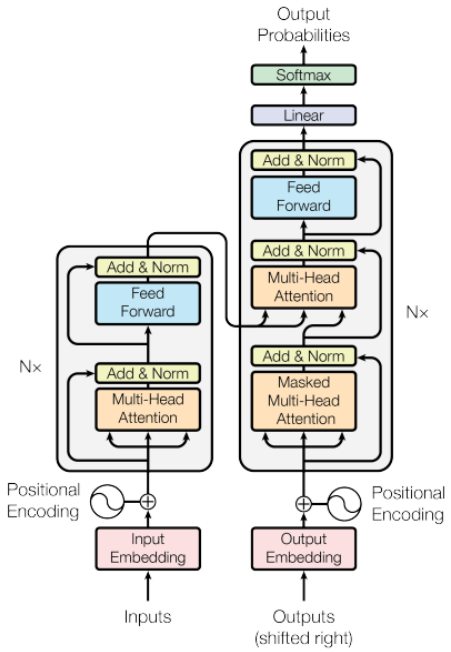

Figure 2: Architecture inspired by Vaswani et al. [4]

Figure 2 details the model's internal architecture. The use of a decoder based on attention mechanisms allows for the establishment of global dependencies between elements in the sequence, unlike recurrent models, which rely on the sequential propagation of information. This property is particularly well-suited to modeling sign languages, which are characterized by simultaneous interactions between multiple gestural components.

3.1. Data Representation

Gesture generation is formulated as a problem of predicting sequences of poses [8]. Each gesture is represented as a temporal sequence of articulatory keypoints extracted from sign language videos.

At each time step t, the pose is described by a set of $K = 75$ keypoints corresponding to the body and hands. Each point is characterized by its spatial coordinates and a visibility index, which allows us to

account for detection uncertainties. Thus, a complete sequence is represented as a tensor of dimension $(T, 75, 3)$, where T denotes the number of frames.

To reduce inter-individual variability and facilitate learning, the coordinates are normalized. Centering is performed relative to the shoulders, and scaling is applied based on the distance between the shoulders. This normalization yields representations that are invariant to variations in position and size within the image.

In addition, the sequences are standardized in length through truncation or padding, accompanied by a temporal mask that identifies the valid frames. This step is essential to ensure compatibility with sequential models and attention mechanisms.

3.2. Architecture of the Proposed Model

The proposed model is based on an encoder-decoder architecture designed to transform a text sequence into a sequence of articulatory poses. As illustrated in Figure 3, it consists of two main modules: a **linguistic encoder** and a **Transformer-style gestural decoder**, connected by an intermediate contextual representation.

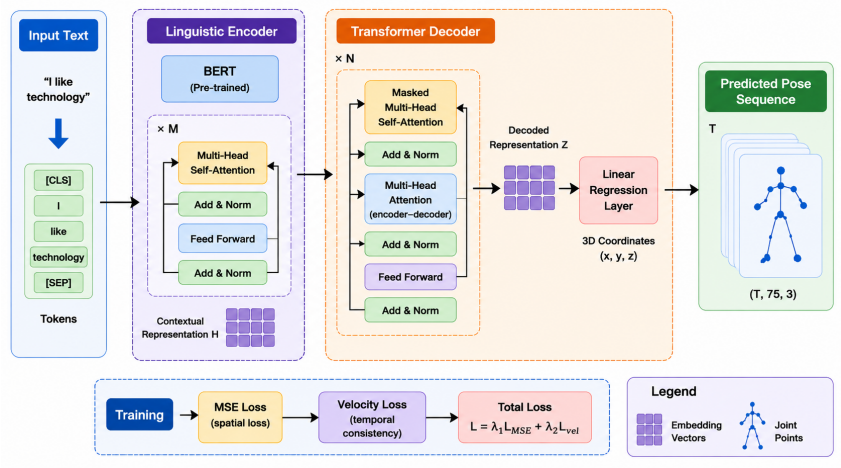

Figure 3: Architecture of the proposed Transformer-based gesture generation model. Contextual embeddings extracted from DistilBERT are projected into the latent space of the decoder and used to generate temporal sequences of articulatory keypoints.

Language encoder. The input text is encoded using the pre-trained **DistilBERT** [10] model. The text is first segmented into tokens (`[CLS]`, words, `[SEP]`), then processed through 6 successive layers, 12 attention heads, and an embedding dimension of 768. Each block of *Multi-Head Self-Attention* and a *Feed Forward* network is followed by an *Add & Norm* operation. At the output, the encoder produces a contextual representation H in the form of embedding vectors, capturing the semantic and syntactic relationships between the words in the sentence.

Transformer Decoder. The decoder is based on a Transformer architecture [4] comprising 4 identical layers, each featuring an 8-head multi-head attention mechanism and a 256-dimensional feed-forward network. A final linear regression layer is used to predict the three-dimensional coordinates of the articulatory points at each time step. Unlike recurrent architectures [9, 11], it processes data in a parallel manner thanks to its **multi-head attention mechanism**, allowing direct access to the entire sequential context during the generation of poses. This choice is particularly justified in the context of sign languages, where multiple gestural components can interact simultaneously within the same sequence.

Each layer of the decoder performs three operations:

1. **Masked Multi-Head Self-Attention**: ensures the internal consistency of the generated sequence by limiting access to future positions;
2. **Multi-Head Attention (with H)**: performs *cross-attention* between the decoder's queries and the contextual representation H from the encoder, thereby aligning the generated gestures with the linguistic content;
3. **Feed Forward + Add & Norm**: applies a nonlinear transformation position by position to enrich the intermediate representations.

Pose generation is based on **learned temporal queries**, each corresponding to a time step t, enriched by a positional encoding that ensures the temporal order is preserved. The decoded representation Z is finally projected by a **linear regression layer** onto the three-dimensional coordinates (x, y, z) of the 75 articulatory keypoints, producing a sequence of poses of dimension $(T,\ 75,\ 3)$.

Loss function. As shown in the lower part of Figure 2, learning is guided by a composite loss function that combines two complementary terms:

$$L = \lambda_1 \, L_{\text{MSE}} + \lambda_2 \, L_{\text{vel}} \ \ with \ \lambda_1 = 0.8, \lambda_2 = 0.2 \tag{1}$$

where L_{MSE} denotes the **mean squared error** calculated between the positions of the predicted keypoints and those of the reference sequence (spatial accuracy), and L_{vel} denotes a **velocity loss** that penalizes abrupt changes between successive frames (smoothness constraint). The coefficients λ_1 and λ_2 allow for balancing the two objectives during optimization. A temporal mask is applied so that only valid frames are taken into account in the loss calculation.

3.3. Learning function

Model training relies on a loss function designed to ensure both the spatial accuracy of gestures and their temporal consistency.

The first component of the loss function is a mean squared error (MSE) calculated between the positions of the predicted keypoints and those of the reference sequence. This metric allows us to evaluate the accuracy of articulatory configurations at each time step.

However, good spatial accuracy does not necessarily guarantee smooth movement. For this reason, a second component is introduced, based on the difference in velocity between successive frames. This additional loss term aims to smooth the temporal dynamics of the gestures by limiting abrupt variations and jerkiness.

The overall loss function thus combines these two terms:

- a **mean squared error (MSE)** on the positions of the keypoints, ensuring spatial accuracy;
- a **velocity loss** that penalizes abrupt changes between successive frames, promoting temporal smoothness.

The main component of the loss is defined by:

$$\mathcal{L}_{MSE} = \frac{1}{T} \sum_{t=1}^{T} \|P_t - \hat{P}_t\|^2$$

where P_t represents the actual pose and $\hat{P}_t$ the predicted pose at time t.

In the experiments conducted, the coefficient associated with velocity loss was set to $\lambda_{vel} = 0.2$. This value was chosen to promote the temporal consistency of the generated motions while maintaining a priority on spatial accuracy ensured by the MSE loss.

The velocity loss (L_{vel}) aims to preserve the temporal consistency of the generated movements by penalizing velocity differences between the predicted sequences and the reference sequences. It is defined as:

$$L_{vel} = \frac{1}{T-1} \sum_{t=2}^{T} \left\| (P_t - P_{t-1}) - (\hat{P}_t - \hat{P}_{t-1}) \right\|^2$$

where P_t represents the actual articulatory coordinates at time t, $\hat{P}_t$ represents the coordinates predicted by the model, and T represents the sequence length.

This component helps limit abrupt changes between successive frames and promotes the generation of smoother, more natural movements.

To improve the temporal coherence of the generated gestures, an additional constraint is imposed on the variations in speed between successive frames. This learning framework enables the model to produce gesture sequences that are both precise and naturally structured over time.

To evaluate the effectiveness of the proposed approach, we implement a rigorous experimental protocol, presented in the following section.

4. Experimental protocol

The goal is to ensure a rigorous and reproducible evaluation by specifying the data used, the comparison models, and the metrics selected.

4.1. Data

The experiments are based on a combination of public datasets and locally collected data in order to cover a variety of gesture configurations.

The How2Sign corpus [7] provides multimodal sequences that align text, video, and sign language information in American Sign Language. The WLASL dataset [6] is also used to introduce greater lexical and gestural diversity. In addition to the public corpora, a local corpus was created to evaluate the model's performance under recording conditions that more closely resemble a real-world context. This corpus comprises 500 videos corresponding to approximately 100 signs and expressions commonly used in American Sign Language (ASL). The data were collected from 10 different signers to introduce inter-individual variability in terms of signing style, execution speed, and expressiveness.

The recordings were made using a digital camera under semi-controlled conditions, with particular attention paid to the visibility of the joints of the body and hands. The total duration of the corpus is estimated at approximately 20 hours of footage. The collected videos cover various gestural configurations and multiple repetitions of each sign, thereby enhancing the diversity of the data used for training and evaluation.

After acquisition, the video sequences were processed using MediaPipe Holistic to automatically extract the three-dimensional coordinates of the key articulation points of the body and hands. The extracted data were then verified, standardized, and converted into time series compatible with the proposed model.

This additional corpus allows for the inclusion of further variations related to signing styles, recording conditions, and the movements of the signers.

Each video is converted into a sequence of key articulatory points using a pose estimation module. The data is then organized into temporal tensors, which are compatible with the proposed model.

To ensure a reliable evaluation, the data is divided into three distinct sets: training, validation, and test, in a ratio of 80%, 10%, and 10%. Where possible, the separation is performed with consideration for the signers, in order to prevent any leakage of information between the sets and to better assess the model's generalization ability.

4.2. Model comparison

To evaluate the contribution of the proposed architecture, a comparison is conducted with classical sequential models widely used for modeling temporal data.

Two recurrent architectures are considered as baselines: the **GRU** [9] and the **LSTM** [11]. These models allow for the capture of temporal dependencies through internal memory mechanisms, but remain limited in modeling long-term relationships.

The main model is based on a **Transformer** [4] decoder, which uses an attention mechanism to establish global relationships within sequences. The three models are configured to have comparable capabilities, particularly in terms of latent dimension, to ensure a fair comparison.

This configuration allows us to isolate the effect of the attention mechanism and evaluate its impact on the quality of the generated sequences.

4.3. Evaluation metrics

Performance evaluation is based on several complementary metrics, which allow for the analysis of both the spatial accuracy and the temporal consistency of the generated gestures.

The first metric used is the mean squared error (**MSE**) [8], calculated on the positions of the keypoints. It measures the average deviation between the predicted sequences and the reference sequences, and serves as a direct indicator of spatial accuracy.

To account for the temporal dynamics of gestures, the **Dynamic Time Warping (DTW)** [12] distance is also used. This metric allows us to evaluate the similarity between two sequences while accounting for potential temporal shifts, which is particularly relevant in the context of gestures.

When linguistic annotations are available, a measure of **precision** is calculated to evaluate the correspondence between the generated sequences and the expected lexical units. This metric complements the analysis by shedding light on the semantic consistency of the model.

Accuracy is calculated based on the percentage of predicted keypoints whose spatial error remains below a threshold determined experimentally. In this study, a prediction is considered correct when the Euclidean distance between the predicted keypoint and the reference keypoint remains below a tolerance threshold defined on the normalized coordinates.

Finally, a **qualitative analysis** is performed based on visualizations of articulatory trajectories, allowing for an evaluation of the fluidity and plausibility of the movements beyond the numerical measures.

4.4. System Implementation

The system was implemented in Python using the PyTorch and Hugging Face Transformers libraries. Articulatory keypoints are extracted using a pose estimation system that retrieves the spatial coordinates of the body and hands in three dimensions.

The proposed model is based on an encoder-decoder architecture incorporating a Transformer responsible for generating pose sequences from textual representations. The linguistic representations are obtained from a pre-trained BERT-type encoder, used to extract contextualized semantic embeddings.

The text encoder used in this study is based on DistilBERT, a lightweight version of BERT that generates rich contextual representations while reducing computational costs. Unlike approaches based on static word representations, DistilBERT takes into account the overall context of the sentence to generate a representation vector tailored to each token.

In our architecture, input sentences are first tokenized and then encoded by DistilBERT to obtain 768-dimensional contextual representations for each token. To ensure compatibility with the generation decoder, these representations are projected into a 256-dimensional latent space using a linear layer. The vectors obtained in this way constitute the memory of the Transformer decoder, which uses them to generate the corresponding sequences of articulatory poses. This strategy allows us to benefit from the semantic richness of the representations produced by DistilBERT while maintaining a relatively compact and efficient generation architecture.

The use of DistilBERT also helps improve the system's robustness against lexical and syntactic variations in the text data, while maintaining a computational complexity that is compatible with the model's training constraints. The data is preprocessed prior to training to ensure the homogeneity of

Table 1
Main model training parameters

Parameter	Value
Optimizer	Adam
Learning rate	10^{-4}
Batch size	8
Number of epochs	100
Dropout	0.1
Model dimension (dmodel)	256
Number of attention heads	8
Loss function	MSE + velocity

Table 2
Quantitative comparison of models on the test set. ↓: smaller is better. ↑: larger is better.

Model	MSE ↓	DTW ↓	Accuracy (%) ↑
GRU [9]	0.02053	**0.340**	51.38
LSTM [11]	0.02667	0.410	38.29
Transformer [4]	**0.01578**	0.365	**70.14**

the sequences. This step includes, in particular, the normalization of spatial coordinates, the temporal resizing of sequences, and the removal of noisy or incomplete frames.

Training was performed using the Adam optimizer with a learning rate set to 10^{-4}. The models were trained for 100 epochs with a batch size of 8. To ensure the reproducibility of the results, a random seed of 42 was used to initialize the experiments.

Several regularization mechanisms are also used to limit overfitting, including dropout and early stopping based on the performance observed on the validation set.

5. Results

This section presents the performance of the proposed model and analyzes the impact of methodological choices on the quality of the generated sequences. The evaluation is based on both quantitative metrics and a comparative analysis with state-of-the-art architectures.

5.1. Quantitative results

The performance of the various models is evaluated on the test set using the metrics described above. Table 2 presents a comparison of the results obtained by the Transformer model and the recurrent architectures.

The results show that the Transformer-based model achieves the best performance in terms of spatial accuracy and temporal consistency. In particular, it achieves a mean squared error (MSE) of **0.01578**, significantly lower than that of the GRU (**0.02053**) and the LSTM (**0.02667**). This improvement reflects the model's greater ability to reproduce the articulatory positions of gestures.

In terms of temporal dynamics, measured by the DTW distance, the Transformer also demonstrates competitive performance (**0.36498**), indicating a good match between the generated trajectories and the reference ones. Although the GRU achieves a slightly lower value on the DTW metric, the generated sequences remain less visually stable and exhibit more temporal irregularities.

Furthermore, the linguistic evaluation highlights a significant advantage of the Transformer, with an accuracy of **70.14 %**, compared to **51.38 %** for the GRU and **38.29 %** for the LSTM. This result confirms that the attention mechanism better captures the contextual relationships between linguistic units and associated gestures.

Overall, these results confirm the superiority of the Transformer over recurrent architectures in the context of generating gesture sequences from text.

5.2. Comparative analysis

Analysis of the results highlights several significant differences between the architectures studied.

Recurrent models, such as GRU and LSTM, demonstrate an ability to model local dependencies, but struggle to capture long-term relationships. This limitation results in an accumulation of errors over time, particularly in long sequences, where the overall coherence of the output is degraded.

Conversely, the Transformer benefits from its global attention mechanism, which allows it to directly access the entire textual context at each stage of generation. This property promotes better structuring of sequences and reduces temporal inconsistencies.

Furthermore, the learning curves observed during training show faster and more stable convergence for the Transformer. Recurrent models, on the other hand, exhibit greater fluctuations, suggesting increased sensitivity to initial conditions and training parameters.

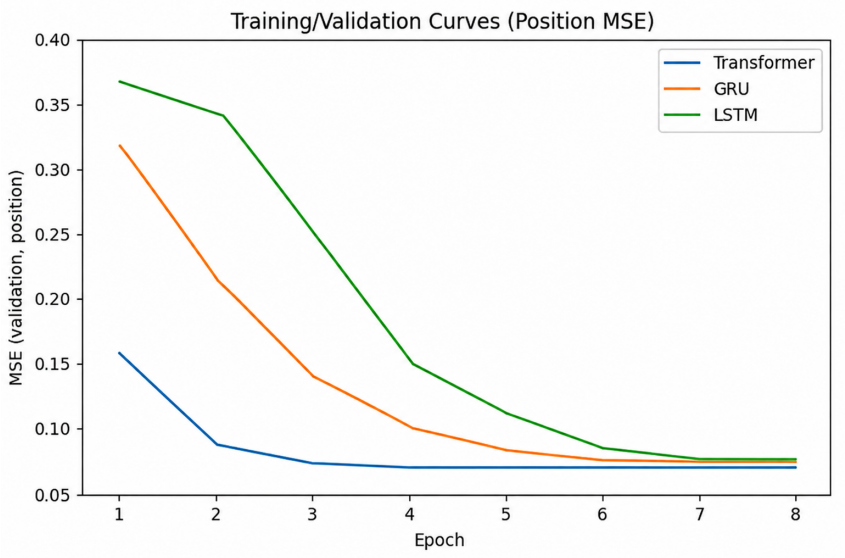

Figure 4: Loss curves on the validation set for the three architectures. The Transformer (blue) converges faster and more stably than the GRU (orange) and the LSTM (green).

These observations confirm that the attention-based architecture provides a more suitable framework for modeling complex multimodal data, such as gestural sequences.

5.3. Impact of methodological choices

Beyond the comparison of architectures, several factors influence the quality of the results obtained.

The introduction of a velocity-based loss component plays an important role in trajectory regularization. By penalizing abrupt changes between successive frames, it helps improve the fluidity of the generated gestures without significantly degrading spatial accuracy.

Similarly, the use of three-dimensional representations of keypoints [8], including a depth or visibility channel, allows for better capture of the spatial structure of gestures. This additional information proves particularly useful in situations involving occlusion or complex movements.

Finally, when glosses are available, using them as linguistic input helps reduce certain semantic ambiguities. However, the proposed model remains capable of operating directly from text, which is a significant advantage in terms of generalization.

Figure 5 illustrates the comparative performance of the different architectures evaluated. The Transformer model achieves the best results with an accuracy of 70.14%, compared to 51.38% for GRU and 38.29% for LSTM.

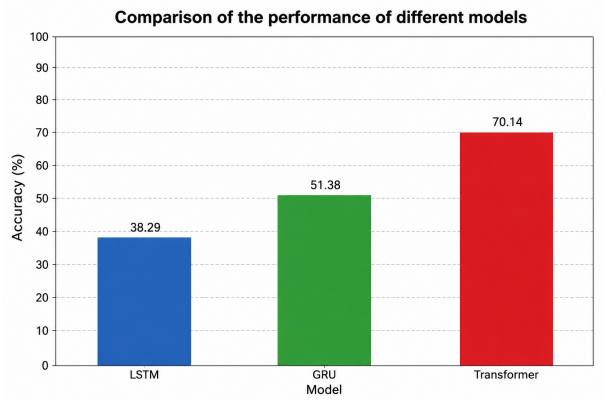

Figure 5: Comparison of the predictive accuracy achieved by the evaluated architectures. The Transformer model obtains the highest accuracy (70.14%), outperforming both GRU (51.38%) and LSTM (38.29%), highlighting its ability to capture long-range temporal dependencies.

This improvement confirms the value of attention mechanisms for modeling complex dependencies between textual representations and gestural trajectories.

5.4. Robustness analysis

To evaluate the model's stability under conditions closer to real-world use, several artificial perturbations were introduced into the test data, including spatial noise, temporal variations, and partial occlusions. The results show that the Transformer architecture maintains relatively stable performance in the face of moderate perturbations. In particular, the model remains capable of producing coherent trajectories despite certain alterations to key points.

Conversely, recurrent architectures exhibit a more rapid decline in performance as the level of perturbation increases. This difference suggests that the attention mechanism contributes to better exploitation of the global context, allowing for the compensation of certain local errors. These observations confirm the relative robustness of the proposed approach under non-ideal conditions.

6. Analysis and Discussion

The results obtained highlight the effectiveness of architectures based on attention mechanisms for the automatic generation of gesture sequences from text. Beyond the quantitative performance metrics observed, an analysis of the generated sequences provides a better understanding of the strengths and limitations of the proposed approach.

The performance results are consistent with trends reported in recent work based on Transformer architectures [5]. Although experimental protocols, datasets, and evaluation metrics differ from one study to another, the results suggest that attention mechanisms offer a better ability to model complex temporal dependencies than classical recurrent architectures. Comparisons with GRU and LSTM models confirm this trend, with the Transformer achieving the best performance in terms of accuracy, temporal consistency, and prediction error.

The qualitative analysis shows that the generated sequences generally exhibit good spatial and temporal coherence. For simple gestures or isolated lexical units, the produced trajectories closely follow the reference movements and maintain a consistent dynamic. In more complex sequences involving rapid transitions or high-amplitude movements, the model maintains satisfactory overall consistency, despite the occasional occurrence of slight articulatory inaccuracies. These deviations remain limited, however, and do not significantly affect the general intelligibility of the generated movements.

One of the main factors explaining this performance lies in the use of the Transformer's attention mechanism [4]. Unlike recurrent architectures, which process information sequentially, the Transformer

simultaneously exploits the entire textual context. This property allows for better consideration of long-range relationships between linguistic elements and contributes to the generation of more stable and coherent gestural sequences.

Furthermore, incorporating a velocity-based loss term into the learning function plays an important role in the quality of the movements produced. By constraining the temporal variations between successive positions, this component promotes the fluidity of articulatory trajectories and reduces abrupt movements that could compromise the realism of the generated gestures.

The experimental comparisons conducted in this study focused primarily on two standard recurrent architectures: LSTM and GRU models. These models were selected due to their frequent use in sequential modeling and motion generation research. It should be noted, however, that several recent approaches to gesture generation also rely on attention-based architectures or advanced variants of Transformers. A direct experimental comparison with all of these methods was not conducted within the scope of this study, primarily due to differences in protocols, corpora, and metrics reported in the literature. A broader comparative evaluation thus represents an important avenue for future research.

Despite the encouraging results obtained, several limitations must be taken into account. The first concerns data availability. Although the public How2Sign and WLASL corpora were supplemented by a local corpus of 500 videos collected from 10 signers, the diversity of the data remains relatively limited given the complexity of sign languages. Individual variations related to signing style, speed of execution, or expressiveness can influence the system's performance and limit its ability to generalize to new users or other usage contexts.

A second limitation lies in the very nature of the representation used. The model relies exclusively on three-dimensional articulatory keypoints and does not account for certain essential components of sign language, notably facial expressions, head movements, and other non-manual cues [2]. Yet these elements play a fundamental role in conveying meaning and linguistic nuances.

Furthermore, despite the performance achieved, the computational cost associated with Transformer architectures remains relatively high, particularly when processing long sequences. This limitation may restrict the system's use in certain contexts that require real-time execution or have limited hardware resources.

Finally, this study does not include a human evaluation of the generated sequences. Performance was measured using quantitative metrics such as mean squared error (MSE), DTW distance, and articulatory point accuracy. Although these indicators provide an objective measure of prediction quality, they do not allow for a full assessment of the naturalness, acceptability, or intelligibility of the gestures from the end-users' perspective. The future involvement of sign language experts as well as deaf or hard-of-hearing individuals will enable a more comprehensive evaluation of the communicative quality of the generated sequences.

Despite these limitations, the results demonstrate the potential of Transformer-based approaches for text-to-sign machine translation [5]. They open up promising avenues for the development of more robust, multimodal, and generalizable systems capable of contributing to improved digital accessibility and inclusive communication.

7. Conclusion

This study explored the problem of automatically generating gesture sequences from text using an architecture that combines DistilBERT and a Transformer decoder. Unlike approaches based on intermediate representations in the form of glosses, the proposed method enables the direct generation of three-dimensional articulatory poses, thereby simplifying the processing chain while reducing dependence on specialized linguistic annotations.

Experiments conducted on the public How2Sign and WLASL corpora, supplemented by a local corpus consisting of 500 videos collected from 10 signers, demonstrated the model's ability to produce coherent and accurate gesture sequences. With an accuracy of 70.14%, superior to that achieved by GRU and LSTM architectures, the results confirm the value of attention mechanisms for modeling the complex

temporal dependencies associated with gestural movements.

Beyond its methodological contributions, this work highlights the potential of artificial intelligence technologies to enhance digital accessibility for deaf and hard-of-hearing individuals. Machine translation systems for sign language could eventually be integrated into educational platforms, digital administrative services, public information systems, and smart urban environments, thereby contributing to the development of more inclusive and accessible digital services.

This research also aligns with a sustainable development perspective by helping to reduce inequalities in access to information and communication technologies. It illustrates how recent advances in artificial intelligence can be harnessed to promote social inclusion and citizen participation.

Future work will focus on expanding the dataset, incorporating non-manual components of sign language such as facial expressions and head movements, and conducting evaluations involving sign language experts and end users. These developments will enable a more precise assessment of the intelligibility, naturalness, and communicative quality of the generated gestures, while enhancing the robustness and generalizability of the proposed models.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] D. Bragg, O. Koller, M. Bellard, L. Berke, P. Boudreault, A. Braffort, N. Caselli, M. Huenerfauth, H. Kacorri, T. Verhoef, C. Vogler, M. R. Morris, Sign language recognition, generation, and translation: An interdisciplinary perspective, in: Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility, ACM, 2019, pp. 16–31. doi:10.1145/3308561.3353774.

[2] O. Koller, Quantitative survey of the state of the art in sign language recognition, 2020. URL: https://arxiv.org/abs/2008.09918, arXiv preprint arXiv:2008.09918.

[3] N. C. Camgöz, S. Hadfield, O. Koller, H. Ney, R. Bowden, Neural sign language translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7784–7793. doi:10.1109/CVPR.2018.00812.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems (NeurIPS), volume 30, 2017. URL: https://arxiv.org/abs/1706.03762.

[5] N. C. Camgöz, O. Koller, S. Hadfield, 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 (CVPR), 2020, pp. 10023–10033. doi:10.1109/CVPR42600.2020.01004.

[6] D. Li, C. Rodriguez, X. Yu, H. Li, Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1459–1469. doi:10.1109/WACV45572.2020.9093512.

[7] A. Duarte, S. Palaskar, L. Ventura, D. Ghadiyaram, K. DeHaan, F. Metze, J. Torres, X. G. i Nieto, How2sign: A large-scale multimodal dataset for continuous american sign language, in: Proceedings of CVPR 2021, 2021, pp. 2735–2744. doi:10.1109/CVPR46437.2021.00276.

[8] J. Zelinka, J. Kanis, Neural sign language synthesis: Words are our glosses, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3492–3501. doi:10.1109/WACV45572.2020.9093516.

[9] K. Cho, B. V. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in: Pro-

ceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1724–1734. doi:10.3115/v1/D14-1179.

[10] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2019.

[11] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (1997) 1735–1780. doi:10.1162/neco.1997.9.8.1735.

[12] S. Salvador, P. Chan, Toward accurate dynamic time warping in linear time and space, Intelligent Data Analysis 11 (2007) 561–580. doi:10.3233/IDA-2007-11506.

[13] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: Proceedings of ICLR 2019, 2019. URL: https://arxiv.org/abs/1711.05101.