Towards Skeletal and Signer Noise Reduction in Sign Language Production via Quaternion-Based Pose Encoding and Contrastive Learning

Guilhem Fauré
guilhem.faure@inria.fr
Université de Lorraine, CNRS, Inria, LORIA
F-54000 Nancy, France

Mostafa Sadeghi
mostafa.sadeghi@inria.fr
Université de Lorraine, CNRS, Inria, LORIA
F-54000 Nancy, France

Sam Bigeard
sam.bigeard@inria.fr
Université de Lorraine, CNRS, Inria, LORIA
F-54000 Nancy, France

Slim Ouni
slim.ouni@loria.fr
Université de Lorraine, CNRS, Inria, LORIA
F-54000 Nancy, France

Abstract

One of the main challenges in neural sign language production (SLP) lies in the high intra-class variability of signs, arising from signer morphology and stylistic variety in the training data. To improve robustness to such variations, we propose two enhancements to the standard Progressive Transformers (PT) architecture (Saunders et al., 2020). First, we encode poses using bone rotations in quaternion space and train with a geodesic loss to improve the accuracy and clarity of angular joint movements. Second, we introduce a contrastive loss to structure decoder embeddings by semantic similarity, using either gloss overlap or SBERT-based sentence similarity, aiming to filter out anatomical and stylistic features that do not convey relevant semantic information.

On the PHOENIX14T dataset, the contrastive loss alone yields a 16% improvement in Probability of Correct Keypoint over the PT baseline. When combined with quaternion-based pose encoding, the model achieves a 6% reduction in Mean Bone Angle Error. These results point to the benefit of incorporating skeletal structure modeling and semantically guided contrastive objectives on sign pose representations into the training of Transformer-based SLP models.

CCS Concepts

• **Computing methodologies** → **Natural language generation**.

Keywords

Sign Language Production, Deep Learning, Contrastive Learning, Pose Encoding

1 Introduction

Sign language is the main way of communication used in the deaf and hard-of-hearing (DHH) community. It leverages a wide range of manual (handshape, location, orientation, movement) and non-manual features (facial expression, body orientation, intensity) to convey ideas through a specific syntax and a rich vocabulary [26, 37, 38].

With around 5% of the global population affected by a disabling hearing loss, and a projection of over 700 million DHH people in 2050 according to the World Health Organization [47], it is essential to reduce the communication gap between deaf and hearing people, notably to avoid social exclusion. As a response to this growing need, a part of the research community has been working on developing new technologies for sign language recognition (SLR), sign language translation (SLT)—sign-to-text—and sign language production (SLP)—text-to-sign—tasks [2, 4, 6–8, 34, 40].

In the effort to develop digital tools that foster communication between deaf and hearing communities, the rise of deep learning has led to significant progress in recent years, particularly in SLR and SLT [3, 4, 7, 22, 30]. Since 2020, an increasing number of studies have focused on generating sign sequences from spoken language [1, 8, 9, 16, 24, 34, 39, 41, 42, 45, 50, 51, 55]. Among these, Saunders et al. [34] introduced the Progressive Transformers (PT) architecture, which has since emerged as a standard baseline in the field.

Despite these advances, SLP systems continue to face several fundamental challenges that hinder their usability in real-world applications. Key obstacles include the high intra-class variability of signs, the significant grammatical divergence between signed and spoken languages, and the scarcity of large-scale, annotated datasets with diverse vocabularies [32]. As a result, the generated outputs often lack the intelligibility, fluency, and naturalness required for deployment in practical communication scenarios.

In this work, we specifically address one of the core limitations of current SLP models: the visual variability of sign realizations, which introduces noise during training and impairs generalization. This variability arises primarily from two sources:

- Inter-signers morphological differences, such as variations in bone lengths, which are not fully addressed by standard normalization techniques (e.g. (Stoll et al. [40]));

- Stylistic variations in the performance of a given sign or sentence—manifested through differences in amplitude, velocity, or positional noise—both across signers and within the same signer.

To mitigate the impact of these factors, we build upon the PT architecture and introduce two main contributions:

- We represent skeletal poses using bone rotations encoded as quaternions rather than traditional 3D Cartesian joint coordinates, and replace the mean squared error (MSE) loss with a geodesic loss defined in quaternion space;
- We incorporate a contrastive loss into the training objective to structure the decoder's multi-head self-attention embeddings by pulling closer sequences with similar semantics and pushing apart dissimilar ones. We investigate two variants of this loss: one based on lexical overlap in the associated glosses (similar to the loss used in (Walsh et al. [45]) for the construction of their codebook), and another leveraging sentence Transformer embeddings (SBERT [33]) similarity scores between associated sentences, with the aim to capture subtler semantic relations.

These contributions aim to reduce the effect of non-semantic variability in training data and improve the semantic consistency and expressiveness of generated sign sequences.

We evaluate our approach on the widely used PHOENIX14T dataset. Code and demos are available online[1].

2 Related Work

2.1 Sign Language Production

Early approaches to SLP were primarily based on synthetic animation techniques relying on avatars and lookup tables containing pre-generated sequences for predefined sentences [2, 6, 12, 20, 56]. These methods required the preparation and storage of a large set of sentence-sign pairs, making them costly and limiting their flexibility. Furthermore, the resulting avatar animations were often poorly received by the Deaf community due to their under-articulated, robotic, or unnatural movements [25].

In recent years, progress in deep neural architectures has significantly advanced research in SLP. Stoll et al. [39] were the first to propose generating sign language pose videos from text using a three-stage pipeline: text-to-gloss conversion via a sequence-to-sequence model, motion graph-based sign stitching, and skeletal pose-to-video synthesis using a generative adversarial network (GAN). Saunders et al. [34] introduced a more streamlined autoregressive model that directly maps sentences or glosses to 3D skeletal poses using a Progressive Transformers architecture. They encode each frame's temporal position in the sequence by appending a normalized counter value $\frac{t}{T}$ to its joint embedding. Subsequent extensions of this model have incorporated data augmentation techniques (e.g., adding Gaussian noise, predicting multiple frames simultaneously), adversarial training, and mixture density networks [35], as well as skeletal graph self-attention mechanisms in the decoder [36]. These improvements target the regression-to-the-mean effect in predicted signs and reduce error propagation during decoding. More recent approaches combine Transformer or diffusion-based

architectures with vector quantization techniques to discretize the sign pose space [45, 50, 55]. Typically, this involves a two-stage process: first, a Vector Quantized Variational Autoencoder (VQ-VAE) is trained to encode sequences of sign poses into discrete tokens by constructing a codebook; second, a model is trained to predict these tokens from textual input.

While some recent models aim to generate avatar-based outputs [1, 55], the majority represent sign poses as 2D or 3D skeletal data and optimize a loss function based on the Cartesian coordinates of joints. However, this approach introduces several limitations:

- The same sign performed by individuals with different body morphologies can lead to significant variations in joint coordinates;
- Computing the MSE over all joints tends to underweight the hands—critical for sign articulation—leading to reduced expressiveness;
- This representation ignores the underlying skeletal structure, requiring the model to implicitly learn limb dependencies, which may result in suboptimal performance.

To address these issues, some studies have introduced body-part-specific loss functions [1, 43, 55], or additional representations using bone orientation vectors in $\mathbb{R}^3$, minimizing an MSE between predicted and reference orientations [41]. However, the latter overlooks the non-Euclidian geometry of rotational space, and may misrepresent angular differences, limiting the precision needed to model fine-grained articulations.

Finally, although pose tokenization effectively reduces stylistic variability [45], learning a robust codebook remains a non-trivial challenge, and may constrain the model's ability to generate novel or unseen signs.

2.2 Rotational Pose Encoding

Instead of representing human motion as sequences of joint positions, an alternative is to describe it through bone rotations. In this framework, each pose is reconstructed by recursively applying a sequence of bone rotations to a predefined skeletal structure in a resting ("T") pose, starting from the root joint. This representation helps prevent prediction errors caused by inconsistent bone lengths or anatomically implausible motions.

Rotational pose encoding relative to a given skeletal structure has been used in various works on human motion recognition and prediction [10, 28, 29, 44]. Rotations can be parameterized in several ways, including 3×3 rotation matrices, Euler angles, exponential maps, and quaternions [13, 15]. However, many of these parameterizations present disadvantages for deep learning applications. For example, rotation matrices require enforcing six nonlinear constraints to remain within the 3D rotation group $SO(3)$, while Euler angles are prone to *gimbal lock* when two rotation axes become aligned, leading to the loss of one degree of freedom. More broadly, since $\mathbb{R}^3$ cannot be smoothly mapped to $SO(3)$, exponential maps may also lead to singularities.

Unit quaternions offer a robust and efficient alternative by representing rotations in 4D space, avoiding these common pitfalls. They are numerically stable, support smooth interpolation, and simplify the composition of rotations [13]. Quaternions have been

---

[1]https://github.com/GFaure9/ContQuat-PT

successfully employed in recurrent models for human motion understanding [28], and more recently in sign language processing to construct sign language action embeddings [46].

Despite their strengths, quaternions—like all representations in four or fewer dimensions—are inherently discontinuous representations of 3D rotations, as shown in (Zhou et al. [54]). The authors demonstrate that continuous representations of 3D rotations can be defined in 5D and 6D, making them better suited for learning. However, quaternions remain an attractive choice for our application, due to their compact 4D representation and low computational overhead while resolving common issues of classical 3D rotation representations.

2.3 Contrastive Learning

Contrastive learning is a machine learning paradigm in which models learn more effective representations by comparing samples—pulling positive pairs (similar samples) closer together in the embedding space, while pushing negative pairs (dissimilar samples) further apart [14, 17, 21]. This approach has proven successful in enhancing language embeddings [11], visual representations [5], and in aligning cross-modal embeddings [31]. In the field of sign language technologies, contrastive learning has been primarily explored in sign-to-text translation frameworks [18, 23, 49, 53]. Ye et al. [49] demonstrate that reducing the density of the sign pose representation space via contrastive learning improves SLT performance. In (Jiang et al. [18]) and (Zhou et al. [53]), contrastive learning is employed for visual-language pretraining by encouraging alignment between visual and textual embeddings when they correspond to matching (*ground truth*, *label*) pairs. Lin et al. [23] supervise the learning of visual feature embeddings using Contrastive Concept Mining (CCM): a method that identifies "anchor words" from batch-level sentences and treats two sign sequences as a positive pair if both contain the same anchor word. A similar technique is adopted in (Walsh et al. [45]) for codebook training, where positive and negative pairs are constructed based on gloss overlap.

Our approach draws inspiration from these works but applies contrastive losses directly within the latent space of the decoder's self-attention layers in the PT architecture. We hypothesize that aligning these latent representations to the underlying semantic distribution of the text before cross-modal attention encourages more efficient learning, by filtering out visual features that are not semantically relevant. This aligns with the motivation of (Walsh et al. [45]), which seeks to reduce signer-specific variability and promote person-invariant representations in sign language generation models.

3 Methodology

3.1 Overview

We adopt the PT architecture of Saunders et al. [34] as our backbone, as it is a widely used baseline in SLP and offers a complete, publicly available implementation[2].

As shown in Figure 1, we propose two extensions: (1) pose sequences are encoded via bone rotations using unit quaternions, replacing MSE loss on Cartesian coordinates with a loss based on the more natural geodesic norm; (2) we explore two contrastive
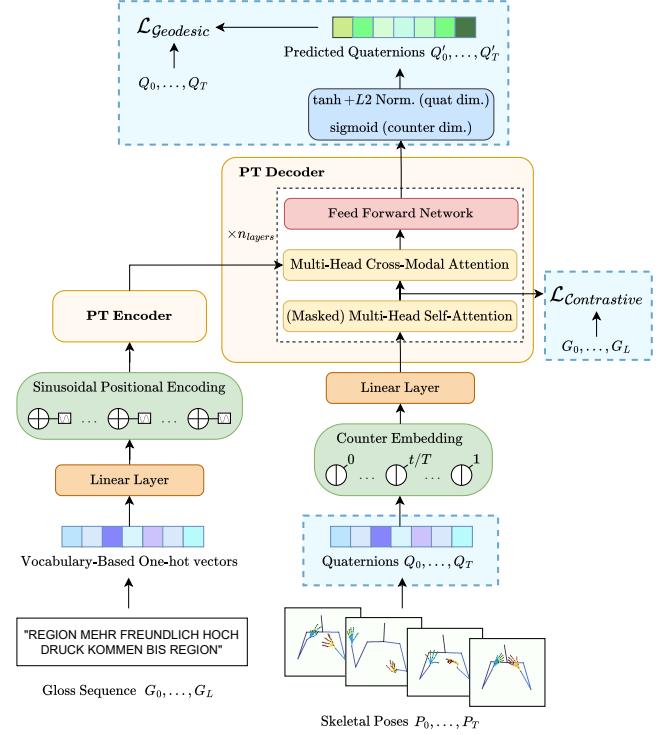
Figure 1: PT model architecture integrating quaternion-based pose encoding and supervised contrastive loss. Blue dotted boxes indicate the modules specific to our contributions.

objectives, supervised by textual input, applied to the decoder's self-attention latent space to guide its structure to reflect semantic relationships. The first variant follows the loss formulation of [45], defining positive and negative pairs based on shared gloss presence. The second aligns similarity matrices between latent features and SBERT sentence embeddings.

3.2 Quaternion-Based Representation of Skeletal Poses

From a skeletal pose in 3D Cartesian coordinates $P := (X_1, ..., X_{N_{joints}})$ lying in $\mathbb{R}^{N_{joints} \times 3}$, given the graph structure of the skeleton and a reference T-pose P_0, we compute the 3D rotation in quaternions representation of the i-th bone as follows:

$$q_i := \left(\cos(\theta_i/2), \sin(\theta_i/2)u^{(i)}\right) \in [-1, 1]^4 \qquad (1)$$

$$\text{where } \begin{cases} \theta_i = \arccos(v^{(i)} \cdot v_0^{(i)}) \\ v^{(i)} = \frac{X_{\text{Child}_i} - X_{\text{Parent}_i}}{\|X_{\text{Child}_i} - X_{\text{Parent}_i}\|}, \quad v_0^{(i)} = \frac{X_{\text{Child}_i}^0 - X_{\text{Parent}_i}^0}{\|X_{\text{Child}_i}^0 - X_{\text{Parent}_i}^0\|} \\ u^{(i)} = \frac{v^{(i)} \times v_0^{(i)}}{\|v^{(i)} \times v_0^{(i)}\|} \quad (\text{'}\times\text{': classical cross product}) \end{cases}$$

Hence, for each sequence of poses $\mathbf{Y} := (P_0, ..., P_T)$, we obtain the corresponding sequence of bones rotations $\mathbf{R} := (Q_0, ..., Q_T)$ where $Q_t := (q_1[t], ..., q_{N_{bones}}[t])$.

Based on the definition of the geodesic distance between unit quaternions, we define the loss function between predicted rotations $\mathbf{R}'$ and ground truth rotations $\mathbf{R}$ as:
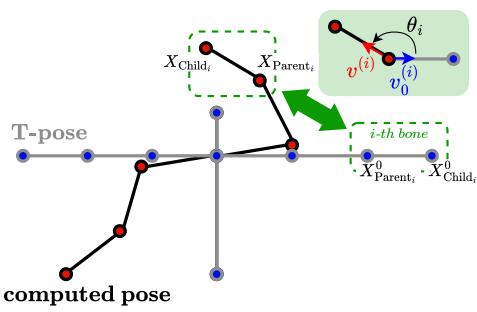
Figure 2: Illustration of bone rotation angle relative to a reference T-pose.

$$\mathcal{L}_{\text{Geo}} := \frac{1}{(T+1)N_{\text{bones}}} \sum_{t=0}^{T} \sum_{i=1}^{N_{\text{bones}}} \arccos\left(2(q_i'[t] \cdot q_i[t])^2 - 1\right) \quad (2)$$

This loss temporally averages the mean rotation angle between predicted and ground truth bone orientations.

Additionally, to enable reconstruction of predicted skeletal pose sequences in 3D Cartesian coordinates by recursively applying rotations from the root joint, we also predict the head node's position by minimizing the following MSE during training:

$$\mathcal{L}_{\text{Root}} := \frac{1}{T+1} \sum_{t=0}^{T} \|X_{\text{Root}}'[t] - X_{\text{Root}}[t]\|_2^2 \quad (3)$$

In constrast, the PT baseline relies exclusively on an MSE loss over joint positions.

3.3 Contrastive Losses

The proposed contrastive losses are incorporated as regularization terms in the overall training objective, in addition to the standard SLP loss—either the MSE on joint positions or $\mathcal{L}_{\text{Geo}} + \mathcal{L}_{\text{Root}}$. A scaling factor λ balances the contribution of the contrastive loss:

$$\mathcal{L}_{\text{Total}} := \mathcal{L}_{\text{SLP}} + \lambda \mathcal{L}_{\text{Cont}} \quad (4)$$

The two contrastive losses we evaluate are presented in the following subsections.

3.3.1 Supervision with Glosses. To define the supervised contrastive loss based on input gloss sequences, we follow the method proposed in (Walsh et al. [45]), itself inspired by (Khosla et al. [21]). For each batch of decoder's self-attention hidden representations, we first extract the set of all unique gloss tokens (referred to as anchors, indexed by I) appearing in the batch. For each anchor $i \in I$, we identify within the batch the sequence where i occurs most frequently. This sequence will serve as the reference. The remaining sequences are split into *positives* $A(i)$ (those that also contain i) and *negatives* $B(i)$ (those that do not), as illustrated in Figure 3.

Based on this grouping, we define the following contrastive loss over the l-th self-attention layer's output $\mathbf{Z}_{\text{batch}}^l := (z_1, ..., z_N)$, where N is the batch size:

$$\mathcal{L}_{\text{GlossSupCont}}^{(l)} := -\sum_{i \in I} \log\left(\frac{1}{|A(i)|} \sum_{a \in A(i)} \frac{\exp(\frac{z_{f(i)} \cdot z_a}{\tau})}{\sum_{b \in B(i)} \exp(\frac{z_{f(i)} \cdot z_b}{\tau})}\right)$$
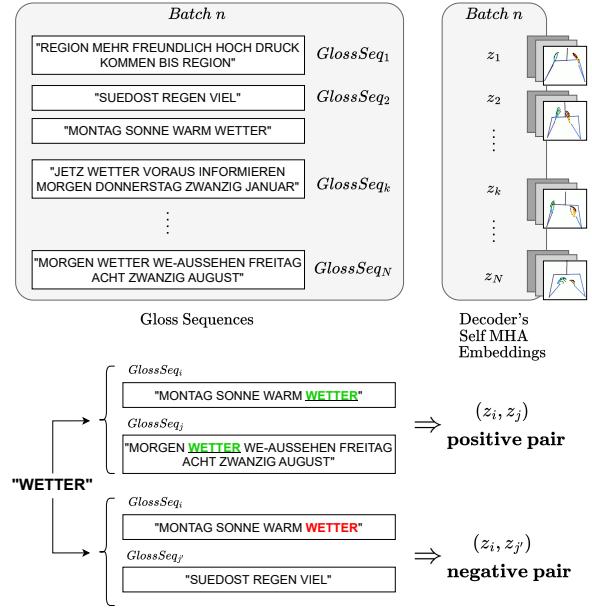$$(5)$$

Figure 3: Definition of positive and negative pairs for the computation of $\mathcal{L}_{\text{GlossSupCont}}$.

$$\text{with} \begin{cases} f(i) := \arg\max_k \{\sum_{m \in k\text{-th Gloss Sequence}} \mathbf{1}_{m=i}\} \\ A(i) := \{a \in [\![1, N]\!] \mid i \in a\text{-th Gloss Sequence}\} \setminus \{f(i)\} \\ B(i) := \{b \in [\![1, N]\!] \mid i \notin b\text{-th Gloss Sequence}\} \end{cases}$$

We average the per-layer losses to obtain the final objective:

$$\mathcal{L}_{\text{GlossSupCont}} := \frac{1}{n_{\text{layers}}} \sum_{l=1}^{n_{\text{layers}}} \mathcal{L}_{\text{GlossSupCont}}^{(l)} \quad (6)$$

3.3.2 Supervision with SBERT Embeddings. To incorporate finer knowledge of semantic relationships between sequences embeddings, we build an alternative loss based on the cosine similarity between input sentences once embedded via a sentence Transformer (SBERT) [33]. Upstream, the embeddings $(\text{SBERT}_k)_k$ of input sentences are thus computed using the 'all-MiniLM-L6-v2' model from the *Hugging Face* library[3]. These embeddings are of dimension 384. Hence, to match the SBERT embedding size before computing the loss, the outputs of the decoder's self-attention blocks are first averaged along the temporal dimension via average pooling, and then projected through a linear layer (cf. Figure 4).

These previous steps enable the computation of the following loss over the batch output of the l-th self-attention layer:

$$\mathcal{L}_{\text{SBERTSupCont}}^{(l)} := \frac{N(N-1)}{2} \sum_{1 \le i < j \le N} d_{i,j}^2 \quad (7)$$

with $d_{i,j} := \mathbf{sim}(\mathbf{g}(z_i), \mathbf{g}(z_j)) - \mathbf{sim}(\text{SBERT}_i, \text{SBERT}_j)$ (8)

Where $\mathbf{sim}(x, y) := \frac{x \cdot y}{\|x\| \|y\|}$ denotes cosine similarity, and $\mathbf{g}(\cdot)$ is the transformation applied by the projection layers. The goal is to

---

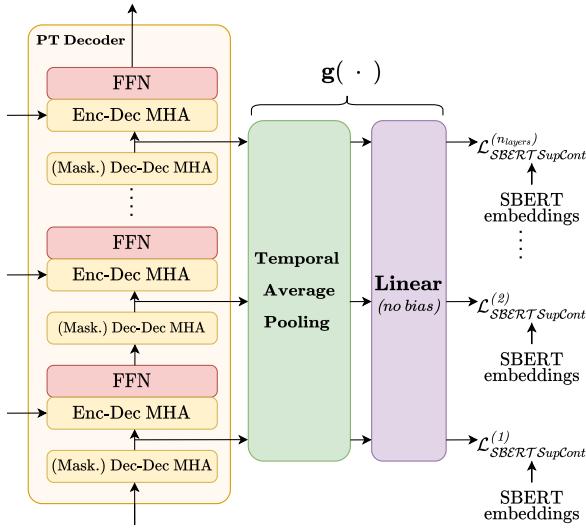[3]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Figure 4: Projection into latent space prior to $\mathcal{L}_{\text{SBERTSupCont}}$ computation. Decoder's self-attention outputs are dimensionally aligned with SBERT embeddings.

align the similarity matrices computed from pose embeddings and SBERT embeddings (see Figure 5), such that the resulting latent spaces are structured according to semantic relationships, while minimizing the influence of non-semantic features.
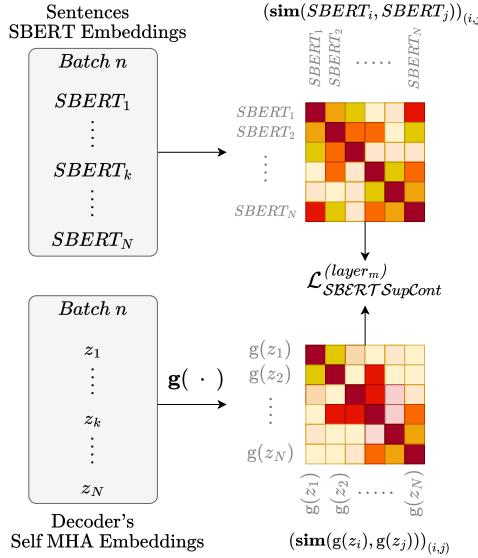

Figure 5: Computation of similarity matrices of SBERT and batch samples $\mathbf{g}(z)$ for $\mathcal{L}_{\text{SBERTSupCont}}$.

As for the first variant of contrastive objective, the overall loss is finally obtained by averaging accross all layers:

$$\mathcal{L}_{\text{SBERTSupCont}} := \frac{1}{n_{\text{layers}}} \sum_{l=1}^{n_{\text{layers}}} \mathcal{L}_{\text{SBERTSupCont}}^{(l)} \qquad (9)$$

4 Experiments

4.1 Experimental Settings

4.1.1 Dataset. We train and evaluate our models on the widely used Phoenix14T dataset, which comprises 8,257 sequences of German Sign Language (DGS) performed by 9 signers, covering 1,066 glosses and a vocabulary of 2,887 unique words [32]. While limited, this dataset is a standard benchmark in SLP, making it a reliable starting point for assessing model performance and comparison with existing methods before scaling to richer datasets.

4.1.2 Preprocessing. 3D skeletal coordinates are extracted using MediaPipe's pose and hand landmarks detection[4]. Joint positions are then refined following the method of Zelinka and Kanis [52], which interpolates missing joints and applies inverse kinematics to correct misplacements while preserving bone length consistency. Finally, skeletons are normalized as in (Stoll et al. [40]), based on shoulder-to-shoulder distance to reduce size variation across subjects.

We use gloss sequences as input for the generation process, following the original PT paper [34] and subsequent works [41, 45, 48].

4.1.3 Evaluation Metrics. We evaluate the tested configurations using standard metrics in SLP to quantify the alignment between generated and reference skeletons. Specifically, we compute the Mean Joint Error (MJE), defined as the Euclidian distance between predicted and ground truth joints averaged over all joints and times steps, as used in prior work [1, 41, 45]. Similarly, we define the Mean Bone Angle Error (MBAE) as the mean angular deviation (in degrees) between predicted and reference bones. It quantifies articulation accuracy independently of bone length, which makes it particularly appropriate for evaluating our quaternion-based variant.

We also compute the Probability of Correct Keypoint (PCK), which measures the proportion of predicted joints falling within a joint-specific neighborhood of their corresponding ground truth positions in the image plane. This neighborhood is defined for each joint, projected onto the (x, y) plane, as a threshold α of the radius of its bounding disk. As in (Kapoor et al. [19]), we choose $\alpha = 0.2$. This metric accounts for varying spatial scales across different body parts.

For both MJE and PCK, sequences are first aligned using Dynamic Time Warping (DTW) based on Euclidean distance between joint coordinates. For MBAE, the applied DTW is instead computed to minimize angular differences between corresponding bones.

It is worth noting that our focus is ultimately on the relative changes in the reported metrics with respect to the PT baseline, rather than on their absolute values. This approach allows us to assess whether the proposed changes lead to measurable improvements within PT-like architectures.

4.1.4 Implementation Details. We retain the original configuration of the PT model from the reference repository, using 2 layers, 4 attention heads, and an embedding size of 512 for both the encoder and decoder. The temperature parameter for the $\mathcal{L}_{\text{GlossSupCont}}$ is set to $\tau = 1$. Moreover, for $\mathcal{L}_{\text{GlossSupCont}}$, we set the scaling factor to $\lambda = 10^{-4}$ and the batch size to 64.

---

[4]https://github.com/google-ai-edge/mediapipe

Training is conducted on an NVIDIA GeForce RTX 2080 Ti GPU. No significant computational overhead is observed between the baseline and the quaternion-based variant, with training times averaging ~1h50 for 1000 epochs. Adding contrastive losses substantially increases runtime, requiring ~3h30 with $\mathcal{L}_{\text{SBERTSupCont}}$ and up to ~11h with $\mathcal{L}_{\text{GlossSupCont}}$.
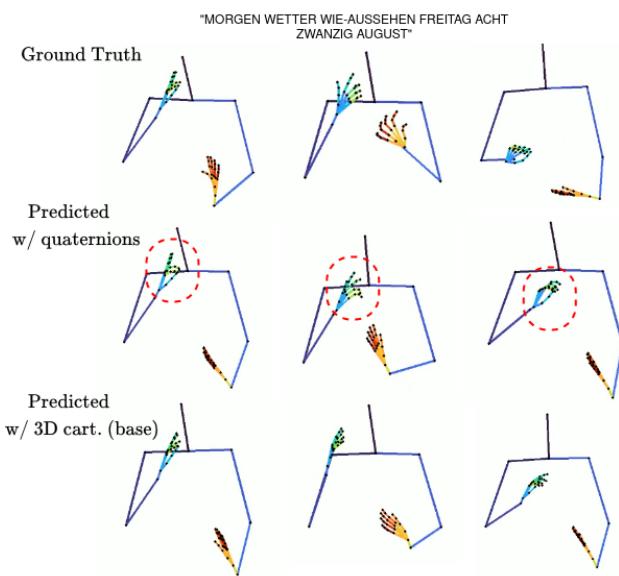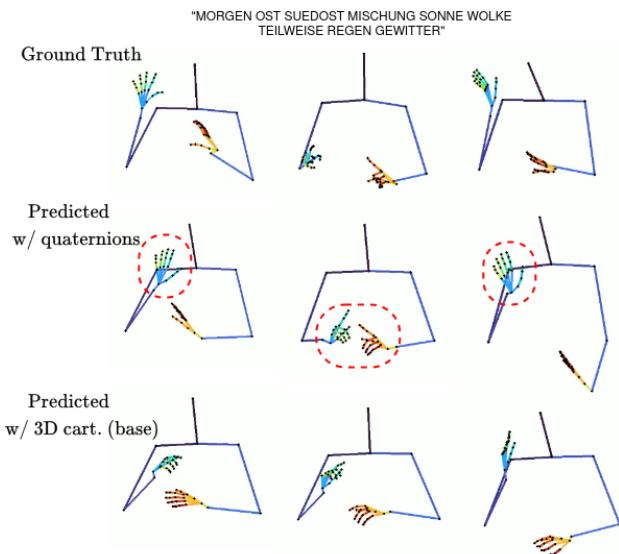
4.2 Results

Figure 6: Qualitative comparison of ground truth and predicted skeletal poses with the base PT model and its quaternion-based variant.

4.2.1 3D Cartesian Positions VS Quaternion-based Rotations. As shown in Table 1, encoding poses through bone rotations using

Table 1: Evaluation metrics on PHOENIX14T test set for different configurations. Values are reported as MEAN$^{\pm\text{STD}}$. A bold score indicates the best result. The second-best result is <u>underlined</u>.

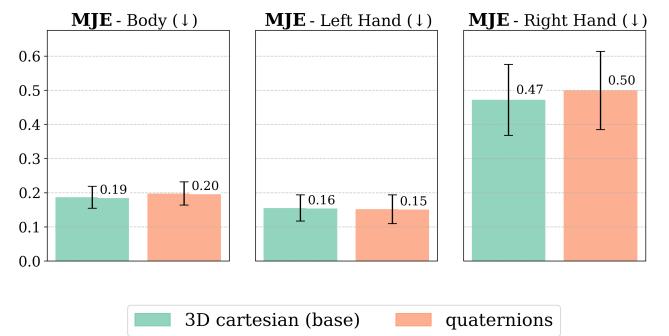|		MJE($\downarrow$)	MBAE($\downarrow$)	PCK($\uparrow$)
3D cart. (base)		$0.41^{\pm0.08}$	$36.93^{\pm6.74}$	$0.25^{\pm0.11}$
w/ gloss cont.		$\mathbf{0.40^{\pm0.08}}$	$36.02^{\pm6.95}$	$\mathbf{0.29^{\pm0.12}}$
	λ			
w/ SBERT cont. batch = 64	*0.0001*	$0.40^{\pm0.08}$	$36.68^{\pm6.97}$	$0.28^{\pm0.12}$
	0.0005	$\underline{0.40^{\pm0.08}}$	$36.57^{\pm6.88}$	$\underline{0.28^{\pm0.12}}$
	0.001	$0.40^{\pm0.08}$	$36.76^{\pm6.76}$	$0.28^{\pm0.13}$
	0.005	$0.41^{\pm0.08}$	$38.14^{\pm6.67}$	$0.26^{\pm0.11}$
	0.01	$0.42^{\pm0.08}$	$38.93^{\pm6.56}$	$0.26^{\pm0.11}$
	0.1	$0.44^{\pm0.09}$	$39.52^{\pm7.14}$	$0.24^{\pm0.11}$
	batch			
w/ SBERT cont. $\lambda = 0.001$	*128*	$0.43^{\pm0.08}$	$39.06^{\pm6.83}$	$0.25^{\pm0.11}$
	256	$0.42^{\pm0.08}$	$38.97^{\pm7.04}$	$0.25^{\pm0.11}$
quaternions		$0.44^{\pm0.08}$	$\underline{35.66^{\pm7.39}}$	$0.22^{\pm0.10}$
w/ gloss cont.		$0.42^{\pm0.09}$	$\mathbf{34.69^{\pm7.19}}$	$0.26^{\pm0.12}$
	λ			
w/ SBERT cont. batch = 64	*0.05*	$0.44^{\pm0.08}$	$36.54^{\pm7.98}$	$0.24^{\pm0.11}$
	0.1	$0.43^{\pm0.09}$	$36.36^{\pm7.23}$	$0.25^{\pm0.12}$
	1	$0.43^{\pm0.08}$	$36.70^{\pm7.08}$	$0.25^{\pm0.11}$

Figure 7: Bar plots of MJE per skeletal part on PHOENIX14T test set between the base PT model and its quaternion-based variant.

quaternions—paired with geodesic loss optimization (see the "quaternions" row)—leads to a slight relative reduction (−3%) in mean angular error compared to the baseline approach using joint 3D Cartesian positions and MSE loss ("3D cart (base)" row). This specific improvement aligns with the objectives of geodesic loss, which better respects the manifold structure of rotations. Qualitative analysis also reveals that the quaternion-based model often produces crisper and more distinct manual articulations, while the baseline tends to generate smoothed, averaged motions (see Figure 6). These results are consistent with findings from (Tang et al. [41]), which emphasize the benefits of modeling bone orientations.

However, using rotations instead of positions leads to diminished performance on standard joint-based metrics, such as MJE, especially noticeable for the dominant right hand (Figure 7). This performance drop is also reflected in lower PCK scores.

4.2.2 PT Model with Contrastive Objectives. Integrating a contrastive loss into the baseline Progressive transformers model consistently improves performance across all evaluated metrics. In particular, the use of $\mathcal{L}_{\text{GlossSupCont}}$ results in a 16% relative improvement in PCK ("w/ gloss cont." in Table 1), from 0.25 to 0.29, while the SBERT embeddings-based variant achieves a 12% relative increase ("w/ SBERT cont." row), from 0.25 to 0.28. These results are in line with previous studies such as (Walsh et al. [45]) and (Zuo et al. [55]) which reduce pose representation space density through vector quantization.

Interestingly, the model trained with $\mathcal{L}_{\text{SBERTSupCont}}$ performs slightly worse than the one using $\mathcal{L}_{\text{GlossSupCont}}$. A likely explanation lies in the nature of the supervisory signals: SBERT-based supervision induces a smoother, more continuous embedding structure by aligning pose sequence similarities with sentence embedding similarities in [0, 1], whereas gloss-based supervision relies on binary similarity labels from shared glosses. This discretization may lead to stronger clustering effects, aiding training convergence and generalization.

It would be beneficial to evaluate these approaches using metrics that better reflect semantic or linguistic intelligibility. Given that stylistic variation still exists in the evaluation data, standard positional metrics may not fully capture the comprehensibility of a sign. A generated sign may remain highly intelligible despite significant deviation from a reference pose, suggesting that current metrics might underestimate improvements that matter most for end-users.

4.2.3 Combining Quaternions Pose Encoding and Contrastive Losses. Combining quaternion-based pose encoding with contrastive training objectives compensates for the positional metric degradation observed when using geodesic loss alone. This combination also leads to further improvements in angular accuracy. Specifically, training the quaternion-based model with $\mathcal{L}_{\text{GlossSupCont}}$ results in an additional 1° reduction in mean angular error, a 4% drop in MJE (from 0.44 to 0.42), and an 18% increase in PCK (from 0.22 to 0.26) relative to the same model trained without contrastive objective.

These findings suggest that integrating rotational encoding with angular-loss objectives and semantic-aware contrastive losses effectively addresses variability in signer morphology and style, ultimately producing clearer and more consistent sign motions.

5 Conclusions

We introduced and explored two complementary strategies to improve the classical Progressive Transformers model for sign language production, focusing on mitigating the impact of morphological and stylistic variability among signers. Our experiments, conducted on the Phoenix14T dataset, demonstrate that: (1) Encoding skeletal poses using bone rotations (quaternions) and optimizing them with a geodesic loss leads to more distinct angular motions, particularly for hand and finger articulations; (2) Augmenting the decoder with a contrastive loss that structures self-attention embeddings yields consistent improvements across all metrics, especially when using shared glosses to define positive sequence

pairs; (3) Combining both methods results in further gains in angular precision while preserving joint position accuracy. These results advocate for the systematic inclusion of skeletal structure and rotation-aware representations, along with semantic-guided contrastive learning, in future SLP model training pipelines.

As future work, we plan to incorporate back-translation metrics, such as BLEU, to more accurately evaluate the semantic intelligibility of generated sign sequences, beyond purely spatial or angular error measures. In addition, we aim to refine the contrastive supervision strategy by leveraging sentence Transformer embeddings to define positive and negative sequence pairs based on semantic similarity thresholds. This approach could offer a hybrid between the current discrete gloss-based method and the continuous SBERT-based formulation, potentially improving the alignment of the learned pose representations with semantic meaning. Finally, given the known limitations of the Phoenix14T dataset, future work will involve evaluating our approach on the more recent Mediapi-RGB French Sign Language dataset, which offers greater diversity and scale, with over 86 hours of video [27].

Acknowledgements

References

[1] Vasileios Baltatzis, Rolandos Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. 2024. Neural Sign Actors: A diffusion model for 3D sign language production from text. 1985–1995. doi:10.1109/CVPR52733.2024.00194

[2] Andrew Bangham, Stephen Cox, R. Elliott, John Glauert, I. Marshall, S. Rankov, and Mariah Wells. 2000. Virtual signing: capture, animation, storage and transmission-an overview of the ViSiCAST project. 6/1 – 6/7. doi:10.1049/ic:20000136

[3] Necati Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural Sign Language Translation. doi:10.1109/CVPR.2018.00812

[4] Necati Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. *Multi-channel Transformers for Multi-articulatory Sign Language Translation*. 301–319. doi:10.1007/978-3-030-66823-5_18

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations *(ICML'20)*. JMLR.org, Article 149, 11 pages.

[6] Stephen Cox, Mike Lincoln, Judy Tryggvason, Mel Nakisa, Mariah Wells, Marcus Tutt, and Sanja Abbott. 2002. TESSA, a system to aid communication with deaf people. *Annual ACM Conference on Assistive Technologies, Proceedings*, 205–212. doi:10.1145/638249.638287

[7] Runpeng Cui, Hu Liu, and Changshui Zhang. 2017. Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition by Staged Optimization. 1610–1618. doi:10.1109/CVPR.2017.175

[8] Sen Fang, Chen Chen, Lei Wang, Ce Zheng, Chunyu Sui, and Yapeng Tian. 2025. SignLLM: Sign Language Production Large Language Models. arXiv:2405.10718 [cs.CV] https://arxiv.org/abs/2405.10718

[9] Sen Fang, Chunyu Sui, Yanghao Zhou, Xuedong Zhang, Hongbin Zhong, Yapeng Tian, and Chen Chen. 2025. SignDiff: Diffusion Model for American Sign Language Production. arXiv:2308.16082 [cs.CV] https://arxiv.org/abs/2308.16082

[10] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. 2015. Recurrent Network Models for Human Dynamics. 4346–4354. doi:10.1109/ICCV.2015.494

[11] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 6894–6910. doi:10.18653/v1/2021.emnlp-main.552

[12] John RW Glauert, Ralph Elliott, Stephen J Cox, Judy Tryggvason, and Mary Sheard. 2006. Vanessa–a system for communication between deaf and hearing people. *Technology and disability* 18, 4 (2006), 207–216.

[13] F. Sebastian Grassia. 1998. Practical parameterization of rotations using the exponential map. *Journal of Graphics Tools* (1998).

[14] R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality Reduction by Learning an Invariant Mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 2. 1735–1742. doi:10.1109/CVPR.2006.100

[15] Fei Han, Brian Reily, William Hoff, and Hao Zhang. 2016. Space-Time Representation of People Based on 3D Skeletal Data: A Review. *Computer Vision and Image Understanding* 158 (01 2016). doi:10.1016/j.cviu.2017.01.011

[16] Jiayi He, Xu Wang, Ruobei Zhang, Shengeng Tang, Yaxiong Wang, and Lechao Cheng. 2025. Text-Driven Diffusion Model for Sign Language Production. doi:10.48550/arXiv.2503.15914

[17] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2021. A Survey on Contrastive Self-Supervised Learning. *Technologies* 9, 1 (2021). doi:10.3390/technologies9010002

[18] Zifan Jiang, Gerard Sant, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling. 2024. SignCLIP: Connecting Text and Sign Language by Contrastive Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 9171–9193. doi:10.18653/v1/2024.emnlp-main.518

[19] Parul Kapoor, Rudrabha Mukhopadhyay, Sindhu Hegde, Vinay Namboodiri, and C.V. Jawahar. 2021. Towards Automatic Speech to Sign Language Generation. 3700–3704. doi:10.21437/Interspeech.2021-1094

[20] Kostas Karpouzis, George Caridakis, S-E Fotinea, and Eleni Efthimiou. 2007. Educational resources and implementation of a Greek sign language synthesis architecture. *Computers & Education* 49, 1 (2007), 54–74.

[21] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 18661–18673. https://proceedings.neurips.cc/paper_files/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf

[22] DONGXU LI, Chenchen Xu, Xin Yu, Kaihao Zhang, Benjamin Swift, Hanna Suominen, and Hongdong Li. 2020. TSPNet: Hierarchical Feature Learning via Temporal Semantic Pyramid for Sign Language Translation. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 12034–12045. https://proceedings.neurips.cc/paper_files/paper/2020/file/8c00dee24c9878fea090ed070b44f1ab-Paper.pdf

[23] Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. 2023. Gloss-Free End-to-End Sign Language Translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 12904–12916. doi:10.18653/v1/2023.acl-long.722

[24] Jian Ma, Wenguan Wang, Yi Yang, and Feng Zheng. 2024. MS2SL: Multimodal Spoken Data-Driven Continuous Sign Language Production. In *ACL*.

[25] Masahiro Mori, Karl F. MacDorman, and Norri Kageki. 2012. The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine* 19, 2 (2012), 98–100. doi:10.1109/MRA.2012.2192811

[26] Adrián Núñez-Marcos, Olatz Perez de Viñaspre, and Gorka Labaka. 2023. A survey on Sign Language machine translation. *Expert Systems with Applications* 213 (2023), 118993. doi:10.1016/j.eswa.2022.118993

[27] Yanis Ouakrim, Hannah Bull, Michèle Gouiffès, Denis Beautemps, Thomas Hueber, and Annelies Braffort. 2024. Mediapi-RGB: Enabling Technological Breakthroughs in French Sign Language (LSF) Research through an Extensive Video-Text Corpus. In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 2: VISAPP*, Vol. 2. Rome, Italy. doi:10.5220/0012372600003660

[28] Dario Pavllo, David Grangier, and Michael Auli. 2018. QuaterNet: A Quaternion-based Recurrent Model for Human Motion. In *British Machine Vision Conference (BMVC)*.

[29] Vladimir Pavlovic, James M Rehg, and John MacCormick. 2000. Learning Switching Linear Models of Human Motion. In *Advances in Neural Information Processing Systems*, T. Leen, T. Dietterich, and V. Tresp (Eds.), Vol. 13. MIT Press. https://proceedings.neurips.cc/paper_files/paper/2000/file/ca460332316d6da84b08b9bcf39b687b-Paper.pdf

[30] Junfu Pu, Wengang Zhou, and Houqiang Li. 2018. Dilated Convolutional Network with Iterative Optimization for Continuous Sign Language Recognition. 885–891. doi:10.24963/ijcai.2018/123

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

[32] Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, and Mohammad Sabokrou. 2021. Sign Language Production: A Review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 3451–3461.

[33] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3982–3992. doi:10.18653/v1/D19-1410

[34] Ben Saunders, Necati Camgoz, and Richard Bowden. 2020. *Progressive Transformers for End-to-End Sign Language Production*. 687–705. doi:10.1007/978-3-030-58621-8_40

[35] Ben Saunders, Necati Camgoz, and Richard Bowden. 2021. Continuous 3D Multi-Channel Sign Language Production via Progressive Transformers and Mixture Density Networks. *International Journal of Computer Vision* 129 (07 2021), 1–23. doi:10.1007/s11263-021-01457-9

[36] Ben Saunders, Necati Camgoz, and Richard Bowden. 2021. Skeletal Graph Self-Attention: Embedding a Skeleton Inductive Bias into Sign Language Production. doi:10.48550/arXiv.2112.05277

[37] William C. Stokoe. 1980. Sign language structure. *Annual Review of Anthropology* 9, 1 (1980), 365–390.

[38] William C. Stokoe, Dorothy C. Casterline, and Carl G. Croneberg. 1976. *A Dictionary of American sign language on linguistic principles*. Linstok Press.

[39] Stephanie Stoll, Necati 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* (04 2020). doi:10.1007/s11263-019-01281-2

[40] Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and R. Bowden. 2018. Sign Language Production using Neural Machine Translation and Generative Adversarial Networks. In *British Machine Vision Conference*. https://api.semanticscholar.org/CorpusID:52288950

[41] Shengeng Tang, Jiayi He, Dan Guo, Yanyan Wei, Feng Li, and Richang Hong. 2025. Sign-IDD: Iconicity Disentangled Diffusion for Sign Language Production. *Proceedings of the AAAI Conference on Artificial Intelligence* 39 (04 2025), 7266–7274. doi:10.1609/aaai.v39i7.32781

[42] Shengeng Tang, Feng Xue, Jingjing Wu, Shuo Wang, and Richang Hong. 2024. Gloss-driven Conditional Diffusion Models for Sign Language Production. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21 (05 2024). doi:10.1145/3663572

[43] Sumeyye Tasyurek, Tugce Kiziltepe, and Hacer Keles. 2025. Disentangle and Regularize: Sign Language Production with Articulator-Based Disentanglement and Channel-Aware Regularization. doi:10.48550/arXiv.2504.06610

[44] Graham W Taylor, Geoffrey E Hinton, and Sam Roweis. 2006. Modeling Human Motion Using Binary Latent Variables. In *Advances in Neural Information Processing Systems*, B. Schölkopf, J. Platt, and T. Hoffman (Eds.), Vol. 19. MIT Press. https://proceedings.neurips.cc/paper_files/paper/2006/file/1091660f3dff84fd648efe31391c5524-Paper.pdf

[45] Harry Walsh, Abolfazl Ravanshad, Mariam Rahmani, and Richard Bowden. 2024. A Data-Driven Representation for Sign Language Production. 1–10. doi:10.1109/FG59268.2024.10581995

[46] Hongli Wen and Yang Xu. 2024. Learning to Score Sign Language with Two-Stage Method. In *Computer Applications*, Haipeng Yu, Chengtao Cai, Lan Huang, Weipeng Jing, Xuebin Chen, Xianhua Song, and Zeguang Lu (Eds.). Springer Nature Singapore, Singapore, 34–50.

[47] World Health Organization. 2025. Deafness and hearing loss. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss. Accessed: 2025-05-22.

[48] Pan Xie, Qipeng Zhang, Peng Taiying, Hao Tang, Yao Du, and Zexian Li. 2024. G2P-DDM: generating sign pose sequence from gloss sequence with discrete diffusion model (*AAAI'24/IAAI'24/EAAI'24*). AAAI Press, Article 693, 9 pages. doi:10.1609/aaai.v38i6.28441

[49] Jinhui Ye, Xing Wang, Wenxiang Jiao, Junwei Liang, and Hui Xiong. 2024. Improving Gloss-free Sign Language Translation by Reducing Representation Density. *NeurIPS*.

[50] Aoxiong Yin, Haoyuan Li, Shen Kai, Siliang Tang, and Yueting Zhuang. 2024. T2S-GPT: Dynamic Vector Quantization for Autoregressive Sign Language Production from Text. 3345–3356. doi:10.18653/v1/2024.acl-long.183

[51] Li Yulong, Yuxuan Zhang, Feilong Tang, Mian Zhou, Zhixiang Lu, Haochen Xue, Yifang Wang, Kang Dang, and Jionglong Su. 2025. Beyond Words: AuralLLM and SignMST-C for Precise Sign Language Production and Bidirectional Accessibility. doi:10.48550/arXiv.2501.00765

[52] Jan Zelinka and Jakub Kanis. 2020. Neural Sign Language Synthesis: Words Are Our Glosses. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 3384–3392. doi:10.1109/WACV45572.2020.9093516

[53] Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. 2023. Gloss-free Sign Language Translation: Improving from Visual-Language Pretraining. 20814–20824. doi:10.1109/ICCV51070.2023.01908

[54] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2019. On the Continuity of Rotation Representations in Neural Networks. 5738–5746. doi:10.1109/CVPR.2019.00589

[55] Ronglai Zuo, Rolandos Potamias, Evangelos Ververas, Jiankang Deng, and Stefanos Zafeiriou. 2024. Signs as Tokens: An Autoregressive Multilingual Sign Language Generator. doi:10.48550/arXiv.2411.17799

[56] Inge Zwitserlood, Margriet Verlinden, Johan Ros, and Sanny Schoot. 2005. Synthetic signing for the deaf: Esign. (01 2005), 1–6. https://www.visicast.cmp.uea.ac.uk/