

Sign Stitching: A Novel Approach to Sign Language Production

Harry Walsh, Ben Saunders, and Richard Bowden

CVSSP, University of Surrey
Guildford, United Kingdom

{harry.walsh, b.saunders, r.bowden}@surrey.ac.uk

Abstract. Sign Language Production (SLP) is a challenging task, given the limited resources available and the inherent diversity within sign data. As a result, previous works have suffered from the problem of regression to the mean, leading to under-articulated and incomprehensible signing. In this paper, we propose using dictionary examples and a learnt codebook of facial expressions to create expressive sign language sequences. However, simply concatenating signs and adding the face creates robotic and unnatural sequences. To address this we present a 7-step approach to effectively stitch sequences together. First, by normalizing each sign into a canonical pose, cropping, and stitching we create a continuous sequence. Then, by applying filtering in the frequency domain and resampling each sign, we create cohesive natural sequences that mimic the prosody found in the original data. We leverage a SignGAN model to map the output to a photo-realistic signer and present a complete Text-to-Sign (T2S) Sign Language Production (SLP) pipeline. Our evaluation demonstrates the effectiveness of the approach, showcasing state-of-the-art performance across all datasets. Finally, a user evaluation shows our approach outperforms the baseline model and is capable of producing realistic sign language sequences.

Keywords: Sign Language Production · Continuous Sequence Synthesis
· Human Pose Generation

1 Introduction

Sign Language Production (SLP) is an essential step in facilitating two-way communication between the Deaf and Hearing communities. Sign language is inherently multi-channelled, with channels performed asynchronously and categorised into manual (hands and body) and non-manual (facial, rhythm, stress and intonation) features. For sign language to be truly understandable, both manual and non-manual features must be present.

Analogous to the tone and rhythm used in spoken language, signed language exhibits prosody. The natural rhythm, stress and intonation that signed languages use to convey additional information [50]. Other phenomena, such as the

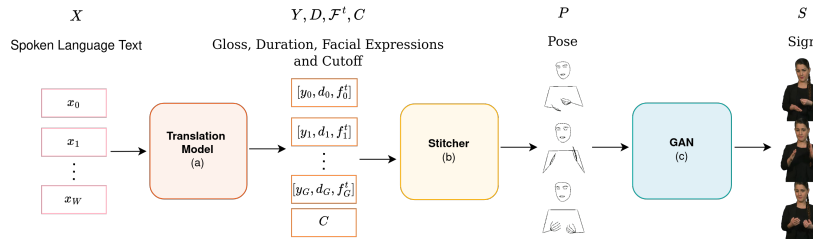


Fig. 1: System overview of our approach to SLP. a) Spoken language to gloss, duration, facial token and cutoff Transformer. b) Stitcher that creates a pose sequence given a list of glosses, facial expressions, durations and a cutoff. c) The SignGAN module that produces a photo-realistic signer conditioned on a pose sequence.

impact of sentiment on the manual and non-manual features have also been studied, showing that emotion can change the form of a sign to exhibit distinctive profiles [35].

Sign language corpora containing linguistic annotation are limited due to the cost and time required to create such annotations [23]. Previous works have attempted to directly regress a sequence of poses from the spoken language or representations such as gloss [14, 15, 36, 39, 52]. However, given that sign language is a low-resource language and the complexity is under-represented in small datasets, previous approaches have suffered from regression to the mean, resulting in under-articulated and incomprehensible signing. Additionally, previous works have implicitly modelled prosody, but due to the limited resources, it is often lost in production.

In this paper, we propose a novel approach to SLP that effectively stitches together dictionary examples to create a meaningful, continuous sequence of sign language. By using isolated signs, we ensure the sequence remains expressive, overcoming previous shortcomings related to regression to the mean. However, each example lacks non-manual features, so we propose a Noise Substitution Vector Quantization (NSVQ) transformer architecture to learn a dictionary of facial expressions that can be added to each sign to create a realistic sequence. To the best of our knowledge, we are the first to explicitly model aspects of signed prosody in the context of SLP. By training a translation model to predict glosses, alongside a duration, facial expression and a cutoff value, we can modify the sequence to eliminate robotic and unnatural movements. Resampling each sign to the predicted duration allows us to alter the velocity associated with signing stress and rhythm [49]. Furthermore, by applying filtering in the frequency domain, we can adjust the trajectory of each sign to create softer signing, akin to how signers modify a sign to convey sentiment [3, 13]. Our approach demonstrates it is capable of modifying the stitched sequence to emulate aspects of prosody seen in the original continuous data. Evaluation of the produced sequence with back translation showcases state-of-the-art performance on all three datasets.

Furthermore, to conduct a realistic user evaluation we use SignGAN, a Generative Adversarial Network (GAN) capable of generating photo-realistic sign language videos from a sequence of poses [37]. Thus, we present a full Text-to-Sign (T2S) SLP pipeline that contains both manual and non-manual features. The user evaluation agrees that the approach outperforms the baseline method [38] and improves the realism of the signed sequence. An overview of the approach can be seen in Fig. 1.

Finally, when ground truth gloss timing is not available, such as on the RWTH-PHOENIX-Weather-2014T (PHOENIX14T) dataset, we leverage our stitching approach to perform sign segmentation. By creating an equivalent sequence and aligning it to the original data, we infer the duration of each sign.

The contributions of this paper can be summarised as:

1. A complete SLP pipeline, capable of generating expressive natural sign language videos from spoken language text.
2. A novel sign stitching approach to effectively join isolated examples and modify them to mimic sign prosody.
3. An approach to create a dictionary of non-manual facial features, that improves the realism of the produced sequence.
4. A novel approach to sign segmentation.
5. Comprehensive evaluation across several metrics and an in-depth user evaluation.

2 Related Work

Sign language Translation: For the last 30 years Computational Sign Language Translation (SLT) has been an active area of research [44]. Initially focusing on isolated Sign Language Recognition (SLR) where a single instance of a sign was classified using a Convolutional Neural Network (CNN) [24]. Subsequent works extended to Continuous Sign Language Recognition (CSLR), which requires both the segmentation of a video into its constituent signs and their respective classification [22]. Later Camgoz et al. introduced the task of Sign-to-Text (S2T) [5] using neural networks, an extension of CSLR that requires the additional task of translation to spoken language. S2T performance was later improved using a Transformer [47]. By using gloss as a supervision on the encoder with a Connectionist Temporal Classification (CTC) loss the model achieved state-of-the-art performance [6]. Although there has been a lot of work since, the architecture has since become the standard when computing back-translation performance [36, 38, 39].

Sign Language Production: SLP is the reverse task to SLT, which aims to translate spoken sentences into continuous sign language. Early approaches to SLP used an animated avatar driven with either motion capture data or parameterised glosses [2, 8–10, 46, 53]. These works all required expensive annotation systems, such as the Hamburg Notation System (HamNoSys) [33] or SigML [18]

and have shown to be unpopular with the Deaf community due to the robotic motion and under articulated signing [34]. None of these approaches attempt to join the isolated signs effectively. Instead opting to play each sign in sequence, with unnatural transitions in between. Furthermore, due to the effect called the ‘uncanny valley’, users do not feel comfortable with the resulting production [28]. Methods that use Motion Capture (MoCap) data achieve better results, but have a limited vocabulary due to the cost of collecting and annotating data [11]. Here, we directly extract skeletons from sign language videos, overcoming the need for expensive MoCap. While Mediapipe has become a standard for this task [26], it sometimes provides incorrect keypoint predictions, negatively impacting downstream tasks. To enhance the quality of pose estimation and uplift the prediction to 3D, Ivashechkin et al. proposed a method that combines Forward Kinematics with a neural network to ensure valid predictions [16]. In this work, we use this additional optimization to extract an accurate sign dictionary.

Research suggests that over 15 million sentence pairs are necessary for Neural Machine Translation (NMT) to outperform statistical approaches [21], but sign Language datasets with parallel text-gloss pairs are restricted to 50,000 sequences [23]. Consequently, sign languages can be classified as a low-resource language, making SLP a challenging task.

Early deep learning SLP approaches used NMT and broke the task down into three steps, Text-to-Gloss (T2G), Gloss-to-Pose (G2P) and Pose-to-Sign (P2S) [43]. Saunders et al. introduced the Progressive Transformer (PT) [38], a transformer architecture that synthesises poses directly from text. Although better results were achieved using gloss as an intermediate representation, the approach suffered from regression to the mean, caused by the lack of training data and the diversity of lexical variants. To reduce the problem, adversarial training and Mixture Density Network (MDN) were applied [36, 39] and since then a range of approaches have been proposed [14, 15, 52]. However, visual inspection of the results shows that the approaches still suffer from regression to the mean, and as a result, they fail to effectively convey the translation. Here we propose a method to effectively join isolated signs, that does not suffer from regression to the mean. By filtering the movement in the frequency domain we ensure the produced sequence is stylistically cohesive. In addition, by modifying the length of each sign we recreate the prosody seen in the original data.

Vector Quantized Models: The first Variational Autoencoders (VAE) model was introduced by Kingma et al. [20], but it struggled to capture fine-grain details. Later the first Vector Quantized Variational Autoencoders (VQ-VAE) architecture was introduced [30]. The model forced the latent space of the VAE to be discrete, and as a result, the model showed state-of-the-art image and audio generation performance. The model uses straight-through gradient estimation to propagate the gradient from the decoder to the encoder. This is necessary as the Vector Quantisation (VQ) operation uses an argmin operation to choose the closest matching codebook entry during training, and hence is non-differentiable. In total, the architecture required three losses to train. Kaiser et al. [17] reduced

the required loss to two by adding an exponential moving average to update the codebook. Recently, Vali et al. [45] introduced the NSVQ VAE, this architecture further reduced the required loss functions to one. The technique approximated the quantization error and substituted it with normalised noise plus the original error. The model is trainable end-to-end, so we apply this approach to learn a dictionary of facial expressions.

Human Pose Synthesis: Research indicates a strong preference among Deaf participants for sign language videos over skeleton representations [48]. Avatar approaches, due to their cartoon-like appearance, are also found to be unpopular, appearing more suitable for children rather than adults [51]. As a result, the field progressed to using photo-realistic approaches. Sanders et al. proposed the SignGAN architecture, a model capable of producing photo-realistic sign language videos, that are comprehensible by a native Deaf signer [37].

The model is designed as a GAN, trained with a combination of 5 losses. Once trained the generator network can synthesise pose and style conditioned images. We apply this model to conduct a user evaluation.

3 Methodology

SLP aims to facilitate the continuous translation from spoken to signed languages by converting a source spoken language sequence, $X = (x_1, x_2, \dots, x_W)$ with W words into a video of photo-realistic sign, denoted as $V = (v_1, v_2, \dots, v_U)$ with U frames. To accomplish this we use two intermediate representations, following Fig. 1 from left to right. First, the spoken language is translated to a sequence of glosses, $Y = (y_1, y_2, \dots, y_G)$, face tokens, $\mathcal{F}^t = (f_1^t, f_2^t, \dots, f_G^t)$ and duration's, $D = (d_1, d_2, \dots, d_G)$, all with length G . Additionally, for each sequence, we predict a low pass cutoff, \mathcal{C} (Fig. 1.a). Each gloss and facial expression is stitched together using these parameters, to produce a continuous sequence of poses, denoted as $P = (p_1, p_2, \dots, p_U)$ with U frames (Fig. 1.b). Finally, we use the pose sequence to condition the SignGAN module allowing us to produce a photo-realistic signer. Next, we provide a detailed explanation of each step in our pipeline, following the order illustrated in Fig. 1 from left to right. We then elaborate on the process of generating the cutoffs, gloss timestamps and the dictionary of facial expressions.

Translation Model Given a spoken language sequence $X = (x_1, x_2, \dots, x_W)$, our goal is to generate a corresponding sequence of glosses $Y = (y_1, y_2, \dots, y_G)$. We design the transformer with four output layers, enabling the model to predict the corresponding duration (in frames) and facial expression for each gloss, plus a low-pass filter cutoff in Hz for each sequence. Thus the model learns the conditional probability $p(Y, D, \mathcal{F}^t, \mathcal{C}|X)$.

The model is an encoder-decoder transformer with Multi-Headed Attention (MHA). The spoken language and gloss sequences are tokenized at the word

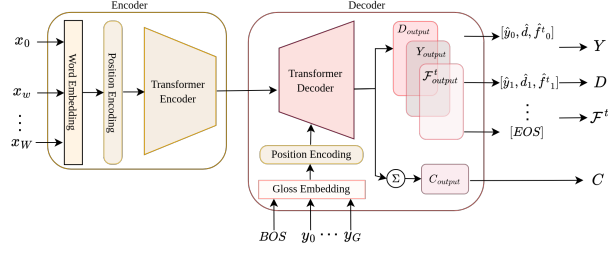


Fig. 2: An overview of the Translation module.

level, and the embedding for a sequence is generated using a token embedding layer. Following the embedding layer, we add sine and cosine positional encoding.

The encoder learns to generate a contextualized representation for each token in the sequence. This representation is then fed into the decoder, which consists of multiple layers of self and cross MHA along with feedforward layers, and residual connections. The gloss, facial expression and duration predictions are obtained by passing the decoder output through their respective output layers. To obtain the cutoff prediction, we pool the decoder embedding across each time step and pass the output through a linear layer. The model is trained end-to-end with the following loss function;

$$L_{total} = \lambda_y \sum_{i=1}^Y \hat{y}_i \log(y_i) + \lambda_f \sum_{i=1}^F \hat{F}_i \log(F_i) + \lambda_d \frac{1}{n} \sum_{i=1}^n (d_i - \hat{d}_i)^2 + \lambda_C (C - \hat{C})^2 \quad (1)$$

Each component is scaled by a factor, λ_y , λ_d , λ_f and λ_C before being combined to give the total loss, L_{total} . The predictions from this model are passed to the stitching module to generate a pose sequence.

Stitching For each dataset, we collect an isolated instance of each gloss in our target vocabulary. For each sign, we extract Mediapipe skeletons [26] and run an additional optimization to uplift the 2D skeletons to 3D [16]. The optimisation uses forward kinematics and a neural network to solve for joint angles, J_a . We choose to store our dictionary as joint angles, as this allows us to apply a canonical skeleton. This ensures the stitched sequence is consistent even if the original signers have different bone lengths. We define a dictionary of, N_s , signs as $DS = [S_1, S_2, \dots, S_{N_s}]$ where each sign in the dictionary consists of a sequence of angles, such that $S_i = (a_1, a_2, \dots, a_{U_s})$ and $a_i \in \mathbb{R}^{J_a}$, where U_s is the duration in frames. In addition we define a learnt dictionary of, N_f , facial expressions as $DF = [F_1, F_2, \dots, F_{N_f}]$, where $F_i \in \mathbb{R}^{U_f \times J \times D}$. As illustrated in Fig. 3, the stitching pipeline is comprised of seven steps, we now detail each step.

Step 1) Given a list of glosses, Y , we select the corresponding signs in our dictionary. If a gloss is absent from the dictionary, we initially lemmatize and format the gloss. If still, we are unable to find a match in the dictionary, we apply

a word embedding model and compute the cosine similarity with all words in the dictionary. We then select the closest sign as the substitute. Such that;

$$j_{sub} = \arg \max_j \left(\frac{\sum_{j=1}^{N_s} \varepsilon(y_q) \cdot \varepsilon(DS_j^y)}{\sqrt{\varepsilon(y_q)^2} \cdot \sqrt{\sum_{j=1}^{N_s} \varepsilon(DS_j^y)^2}} \right) \quad S = DS[j_{sub}] \quad (2)$$

Here $\varepsilon()$ represents the word embedding model, y_q is the query gloss and DS^y is the dictionary's corresponding gloss tags. We find word embeddings capture the meaning of words, enabling substitutions such as replacing RUHRGEBIET (RUHR AREA) with LANDSCHAFT (LANDSCAPE). Simultaneously, in this step, we select the corresponding facial expressions, F , from the dictionary, DF , given the predicted face tokens, \mathcal{F}^t .

Step 2) The selected signs are converted from angles into a 3D canonical pose. We normalise the rotation of the signer, such that the midpoint of the hips is located at the origin and the shoulders are fixed on the y plane. This ensures the skeleton is consistent across all the signs. Consequently, we convert from a sequence of angles $S_n \in \mathbb{R}^{U_s \times J_a}$ to a sequence of poses, $P = (p_1, p_2, \dots, p_{U_s})$ with the same number of frames, U_s . Each pose, p_u , is represented in D -dimensional space and consists of J joints, denoted as $p_u \in \mathbb{R}^{J \times D}$.

Step 3) The dictionary signs often start and end from a rest pose. Therefore, to avoid unnatural transitions we cropped the beginning and end of each sign. For this, we track the keypoint T corresponding to the wrist of the signer's dominant hand and measure the distance travelled.

Thus, for each dictionary sign we create a sequence, $P^\Delta = (p_2^\Delta, p_3^\Delta, \dots, p_{U_s}^\Delta)$ $n \in 2, 3, \dots, U_s$, representing the distance travelled for a dictionary sign. We remove the beginning frames once the sign has moved by a set threshold, α_{crop} . The crop index is given by:

$$\text{index}_{\text{start}} = \arg \max_u \left(\sum_{u=1}^{U_s} P_u^\Delta - \alpha_{crop} \cdot \sum_{u=1}^{\max(U_s)} P_u^\Delta \right) \quad (3)$$

To crop the end, we reverse the order of frames and repeat the process. However, for short sequences, this method may over-crop the sign, rendering it meaning-

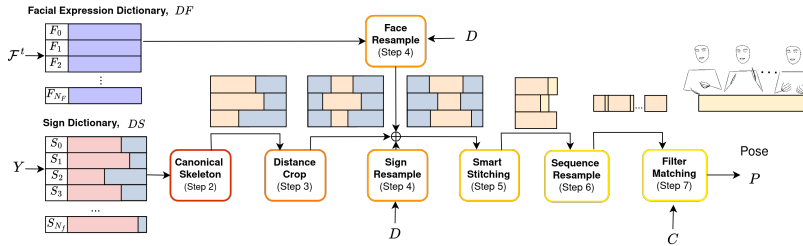


Fig. 3: An overview of the stitching module.

less. We introduce an additional constraint $index_{end} - index_{start} < U_{min}$. If this condition holds, we reduce the threshold, α_{crop} , by a factor of 0.1 and recalculate.

Step 4) As detailed above, we predict the duration of each gloss. Here, we utilize the duration to resample the length of each sign and facial expression, emulating the natural rhythm in the original data. This process involves up-sampling or down-sampling the sign using linear interpolation. However, due to co-articulation in the continuous data, some signs are predicted to be only a few frames long, which may not be sufficient for an isolated example. To address this, we impose a minimum sign length to ensure the produced sequence remains expressive. Once the facial expression and sign are resampled to the same length, we shift and rotate the face onto the body creating the complete skeleton.

Step 5) Having created a list of signs in a canonical pose, cropped, and resampled to match the original data, the next step involves joining these signs into a single sequence using the smart stitching module. The objective is to achieve a natural transition between the end of one sign and the start of the next. To accomplish this, we track the dominant hand of the signer and calculate the distance, Δ , between the end of the first sign and the start of the second sign. Then we can determine the required number of frames, U_{stitch} , needed to create a smooth transition. Such that:

$$U_{stitch} = \arg \min_u (V_{min} < \frac{fps * \Delta}{u} < V_{max}) \quad (4)$$

Where V_{max} and V_{min} are the start and end velocities of the two signs. This calculation ensures that the signer’s velocity is bounded by the end velocity of sign one and the initial velocity of sign two. In cases where multiple solutions exist, we select U_{stitch} that minimizes the standard deviation between the start and end velocity. Following this, we employ linear interpolation to generate the missing frames.

Step 6) We concatenate the signs and the stitched frames to form a single sequence. We then sum all the predicted durations and resample the sequence so it matches the ground truth.

Step 7) Finally, we apply a low-pass Butterworth filter to each keypoint over time [4]. The predicted cutoff value determines which frequencies are removed, and corresponds to the -3dB attenuation point. This step aims to enhance the stylistic cohesiveness of the sequence by smoothing out any sharp, quick movements not present in the original sequence. The transfer function can be formulated as;

$$H(z) = \frac{1}{1 + \left(\frac{z}{e^{j \cdot \omega_c}}\right)^{2N}} \quad (5)$$

Here we apply a 4th order filter thus, N is 4, and the angular cutoff frequency is given by $\omega_c = 2\pi C$. z corresponds to the z -transform of the pose sequence. Applying the bilinear transform to Eq. (5) gives the discrete formula that we apply. This process generates natural pose sequences, that remain expressive and are stylistically cohesive. Next, we map these poses to a photo-realistic signer.

SignGAN Skeleton outputs have been shown to reduce Deaf comprehension compared to a photo-realistic signer [48]. Therefore, to gain valuable feedback from the Deaf community we train a SignGAN model [37].

Given a pose sequence, $P = (p_1, p_2, \dots, p_U)$, generated by our stitching approach the model aims to generate the corresponding video of sign language, $V = (v_1, v_2, \dots, v_U)$ with U frames. Each video frame is conditioned on the appearance of an individual extracted from a style image. For the user evaluation, we use the appearance of a single individual obtained by passing a style image through the style encoder.

Facial Expression Generation To be truly understandable and accepted by the Deaf community non-manual features must be present in the final output. Here we are using a discrete sequence-to-sequence model to generate the translation. Therefore, we must learn a discrete vocabulary of facial expressions, DF , that can be added to the isolated signs. We apply a NSVQ to learn a spatial-temporal dictionary of facial expressions.

For each gloss in the original data, we extract the corresponding face mesh, denoted as F . We then resample each sequence to a constant length, U_f and scale it to be a constant size. Signers in the dataset are often looking off center, therefore we normalize the average direction of the face so that it is looking directly forward. Similar to Fig. 2 (Encoder) we add positional encoding and then embed the sequence using a single linear layer. After the embedding is passed through the transformer encoder to the codebook. The NSVQ codebook learns a set of N_f embeddings. We denote the embedded face sequence and therefore each codebook entry as $F_i^z \in \mathbb{R}^{U_f \times H}$, where H is the embedding dimension. Each input is mapped to one codebook entry, the difference between the selected codebook entry and the input is then simulated using a normally distributed noise source. A product of the simulated noise and the encoder output is then passed to the decoder. We use the counter decoding technique from the PT [38], to drive the decoder. The decoder learns to reconstruct the original face sequence and the input counter values. Thus, the model is trained with the following loss function;

$$L_{Face} = \frac{1}{U_f} \sum_{u=1}^{U_f} ((f_u - \hat{f}_u)^2 + \lambda_{CN}(c_u - \hat{c}_u)^2) \quad (6)$$

Where λ_{CN} is a scaling factor and c is the counter value. Once fully trained we pass each codebook embedding, F_i^z , through the decoder to give the learnt dictionary of facial expressions in Euclidean space, $DF = [F_1, F_2, \dots, F_{N_f}]$.

Duration Generation When ground truth timing information is not available we propose a novel sign segmentation approach based on the stitching method described above. Given the ground truth gloss labels, we generate the stitched sequence, P_{stitch} , but without step 4 (sign resampling).

Comparing the stitch sequence and the ground truth, we find that the motion can vary due to different lexical variants present compared to our dictionary. However, we find that the handshape is often still consistent. So, we take the keypoints that correspond to the signer’s hands and normalise the rotation so that the index finger metacarpal bone is fixed on the y-axis and the palm is fixed on the xy-plane, giving P_{stitch}^H, P^H . Our next step is to align the two sequences so that we can infer the duration of the signs in the ground truth. For this we apply Dynamic Time Warping (DTW), such that;

$$A_{i,j} = DTW(P_{stitch}^H, P^H) \quad (7)$$

As we know the duration of the isolated signs in the stitched sequence, by analysing the alignment path, A_j , we can infer the duration of the signs in the original ground truth sequence.

Cutoff Generation Experiments reveal that each sequence contains a distinct range of frequencies correlated to the signer’s style. Fast motions contain high frequencies, while soft, slow signing involves lower frequencies, typically within the range of 1 to 25 Hz.

To generate the ground truth cutoffs used in training, we once again apply our stitching approach. For each sequence in the ground truth data, P , we produce the equivalent stitched sequence, P_{stitch} . We then apply a low-pass filter to P_{stitch} within the range of 1 to 25 Hz and measure the intersection and set difference of the frequencies, denoted as $(FT(P) \cap FT(P_{stitch}))$ and $(FT(P) \setminus FT(P_{stitch}))$, where the Fourier transform is represented as $FT()$. Subsequently, we fit a parametric spline curve to the intersection and set difference. To determine the cutoff we find the frequency that maximises the intersection while minimising the set difference. This provides the cutoff frequency for that sequence. We opt to use this method over just analyzing the frequency in the original sequence as we do not have an ideal filter. Thus, the butterworth filter has unintended effects on the frequency and phase blow the cutoff.

4 Experimental Setup

Implementation Details In our experiments, we conducted a grid search for optimal hyper-parameters and identified the following settings as the most effective. Our encoder-decoder translation model is constructed with an embedding size of 512 and a feed-forward size of 1024. We find that the optimal number of layers and heads is dataset-dependent, with smaller datasets requiring fewer layers compared to Meine DGS Annotated (mDGS) where we use 3 layers and 4 heads. The models utilise dropout with a probability of 0.1 [41], ReLU activations between layers [1], and pre-layer normalisation for regularisation and training stability. Training employs a ‘reduce on plateau’ scheduler with a patience of 5 and a decrease factor of 0.8. The layers are initialised using a Xavier initializer [12] with zero bias, and during training, Adam optimization is employed [19]. The initial learning rate is set to 10^{-4} , and we train the model until

convergence. During decoding, we utilise a greedy search algorithm. The loss scaling factors, λ_y , λ_d , λ_f and λ_C are set to 1.0, 0.1, 0.3 and 0.2, respectively. When stitching we enforce a minimum sign length, U_{min} of 4 frames. The cropping threshold, α_{crop} is dataset-dependent, we find values between the range of 0.1 to 0.35 most effective. All sequences are subsampled to 12 frames per second (fps) for computational efficiency.

For each dataset, we create a dictionary of 500 facial expressions. We scale the counter loss by 0.01 and we set an embedding dimension of 512. The encoder and decoder are initialized with the same settings as our translation model.

The angular pose representation comprises 104 angles, while the Euclidean representation consists of 61 keypoints (21 for each hand and 19 for the body and face). The face mesh we add includes 128 keypoints, which are a subset of Mediapipe’s 478-face mesh [26].

For comparison, we train a progressive transformer on each dataset until convergence using the parameters from [38].

Datasets The approach is tested on three datasets, the Public Corpus of German Sign Language, 3rd release, the mDGS dataset [23], PHOENIX14T [5] and the BSL Corpus **T** (BSLCPT) [40]. BSLCPT contains 211 participants from 8 regions in the UK, performing 4792 individual signs from a range of age groups. The participants perform narrative, interviews and participate in free conversation. Similarly, mDGS contains 330 participants engaging in free-form signing. Whereas, the PHOENIX14T dataset is extracted from German TV weather broadcasters and contains over 8,000 parallel sequences.

Dictionary In the following experiment, we test two different dictionaries: 1) collected from isolated examples, and 2) a dictionary created from continuous data. Next, we provide further details about each:

Isolated: Here, the signs are sourced from individuals who perform each sign in isolation, typically starting from and returning to a resting position. When experimenting on the BSLCPT we use the Signbank dataset [7], it contains over 3,000 signs and includes all the lexical variants found in the BSLCPT dataset. However no such dictionary exists for the PHOENIX14T and mDGS dataset, therefore we collect a dictionary from a range of sources such as [23]. We find the mDGS dataset has a target vocabulary of 10,801. However, without the gloss variant, we find the core gloss vocabulary reduces to 4,434. We collect a total of 7,206 signs to experiment with. We use the method described in Sec. 3 (Stitching), step 1 to overcome issues with an incomplete vocabulary. To create word embeddings we apply Facebook’s implementation of Fasttext [27]. When experimenting on the BSLCPT we use the English implementation whereas on PHOENIX14T and mDGS we use the German implementation.

Continuous: Here we create a dictionary using the gloss timing annotations. The signs are taken from the test and dev data only, so that the back translation model has not seen the signs during training. As the examples come from the continuous sign, we omit the cropping step of the stitching pipeline. These

dictionaries have an abundance of signs to choose from when stitching. We filter the dictionary and remove short signs as these are most likely co-articulated and therefore not suitable or out of context. We opt to randomly select the first sign in the sequence, and subsequent signs are chosen to ensure the most natural transition. Therefore, we select the sign from the dictionary in which the location of the wrist is closest to the last frame of the previous sign.

Evaluation Metrics To evaluate our approach we employ a CSLR model (Sign Language Transformers [6]) to conduct back-translation, the same as [14, 36, 39, 52]. A model is trained for each dataset and the parameters are frozen so that results are consistent. BLEU [31], Rouge [25], and chrF [32] scores are computed between the predicted text and the ground truth. To evaluate the pose we employ Dynamic Time Warping Mean Joint Error (DTW-MJE), the metric compresses and stretches time to find the best alignment between the ground truth, p , and the predicted pose, \hat{p} .

5 Experiments

5.1 Quantitative Evaluation

Text-to-Gloss Translation Results We start by evaluating the T2G translation performance described in Sec. 3. Tab. 1 shows the performance on all three datasets. We suggest that the difficulty of a dataset is proportional to the vocabulary and the total number of sequences used in training. We find the best performance on PHOENIX14T data which has the highest number of sequences per token, achieving 18.11 BLEU-4. In comparison to previous works, we find by having the model predict duration, face and cutoff we can achieve higher BLEU-1 scores, but at the cost of a lower BLEU-4 in comparison to [38]. On the more challenging mDGS dataset we find a considerably lower BLUE-4 score due to the larger domain of discourse. The BSLCPT has a smaller domain of discourse in comparison to mDGS but has the fewest sequences per token. Thus, understandably we only achieve a BLEU-4 of 1.67 on the test set. Overall we find the model to be performing as expected.

Table 1: The results of translating from Text-to-Gloss on the BSL Corpus T, RWTH-PHOENIX-Weather-2014T and Meine DGS Annotated dataset.

Dataset:	TEST SET						DEV SET					
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	chrF	ROUGE	BLEU-1	BLEU-2	BLEU-3	BLEU-4	chrF	ROUGE
BSLCPT	26.02	11.19	4.15	1.67	23.54	25.06	26.88	11.40	4.72	1.28	23.59	26.96
mDGS	30.13	13.04	5.45	2.36	29.24	31.43	29.72	12.46	4.89	1.87	28.90	30.90
PHOENIX14T	55.48	36.54	25.18	18.11	49.30	53.83	56.55	37.32	25.85	18.74	48.91	54.81
PHOENIX14T [38]	55.18	37.10	26.24	19.10	-	54.55	55.65	38.21	27.36	20.23	-	55.41
PHOENIX14T [42]	50.67	32.25	21.54	15.26	-	48.10	50.15	32.47	22.30	16.34	-	48.42

Text-to-Pose Translation Results Note in the following experiments the back-translation model’s performance (shown as GT top row of Tab. 2, 3 and 4) should be considered the upper limit of performance. In this section, we evaluate the Text-to-Pose (T2P) performance using back translation. To allow for a comparison we train two versions of the PT with the setting presented in [38]. PT is the standard architecture, while PT + GN is trained with Gaussian Noise added to the input. In line with the original paper, we find Gaussian Noise improves the performance, however, our approach still outperforms both models except on DTW-MJE. As discussed previously, other works suffer from regression to the mean caused by the models attempting to minimise their loss function and thus, are incentivised to predict a mean pose. This metric fails to evaluate the content of the sequence, but the higher score does indicate our model is expressive as it is producing sequences further from the mean. For back-translation, we outperform the PT on all metrics. Showing significant improvements in BLEU-1 score of 98% and 269% on the mDGS and BSLCPT dev set (comparing PT + GN and Stitcher (continuous), Tab. 2 and 3).

Deep learning models exhibit a bias toward the data they were trained on and often show poor out-of-domain performance. Unsurprisingly, the performance improves when using the continuous dictionary. We find only a small increase in BLEU-1 of 0.13 on the BSLCPT dev set (Tab. 2), most likely due to the isolated dictionary containing the lexical variants found in the original data. Whereas we see a larger increase on the mDGS dataset (Tab. 3).

Table 2: The results of translating from Text-to-Pose on the BSL Corpus T dataset.

BSLCPT Approach:	TEST SET								DEV SET							
	DTW-MJE	BLEU-1	BLEU-2	BLEU-3	BLEU-4	chrF	ROUGE		DTW-MJE	BLEU-1	BLEU-2	BLEU-3	BLEU-4	chrF	ROUGE	
GT	0.000	17.3	3.96	1.37	0.54	13.00	21.76		0.000	17.32	3.71	1.08	0.39	13.04	21.89	
PT [38]	0.288	4.40	0.65	0.18	0.00	5.80	8.22		0.292	4.00	0.61	0.10	0.00	5.69	8.02	
PT + GN [38]	0.267	4.96	0.55	0.13	0.00	6.38	8.82		0.258	4.47	0.63	0.09	0.00	6.14	8.89	
Stitcher (Isolated)	0.588	16.37	2.86	0.75	0.28	14.07	20.84		0.592	16.39	2.82	0.58	0.00	13.9	19.55	
Stitcher (continuous)	0.575	16.99	3.65	1.03	0.41	14.32	20.65		0.573	16.52	3.19	0.73	0.00	14.34	20.53	

Table 3: The results of translating from Text-to-Pose on the Meine DGS Annotated (mDGS) dataset.

mDGS Approach:	TEST SET								DEV SET							
	DTW-MJE	BLEU-1	BLEU-2	BLEU-3	BLEU-4	chrF	ROUGE		DTW-MJE	BLEU-1	BLEU-2	BLEU-3	BLEU-4	chrF	ROUGE	
GT	0.000	20.87	5.60	1.89	0.80	17.56	23.78		0.000	20.75	5.43	1.81	0.76	17.63	23.41	
PT [38]	0.229	6.11	0.94	0.21	0.05	8.07	8.36		0.228	6.22	0.98	0.17	0.00	8.23	8.44	
PT + GN [38]	0.2245	7.18	1.48	0.40	0.01	8.46	8.38		0.2241	9.22	1.63	0.38	0.01	8.94	8.57	
Stitcher (Isolated)	0.581	16.63	3.75	0.94	0.22	13.39	21.69		0.592	16.9	3.67	0.95	0.32	13.9	21.34	
Stitcher (Continuous)	0.637	18.64	4.17	1.07	0.39	16.86	21.80		0.637	18.27	4.07	1.19	0.43	16.75	21.25	

Previous work has primarily focused on G2P translation, therefore to facilitate a meaningful comparison we present two versions of the model. First, a

G2P version, where we use the ground truth data and just apply the stitching module, and, secondly our T2P approach (translation then stitching). Results for comparison are provided by [52]. We find our approach outperforms previous work on the BLEU-1 to 2 scores increasing the score by 56% and 19%, respectively (comparing Tab. 4, row 7 and 11). We also find significant improvement in ROUGE and chrF metrics.

Table 4: The results of translating from Gloss-to-Pose (G2P) and Text-to-Pose (T2P) on the RWTH-PHOENIX-Weather-2014T dataset.

PHOENIX14T		DTW-MJE	BLEU-1	BLEU-2	BLEU-3	BLEU-4	chrF	ROUGE
Approach:								
	GT	0.000	32.41	20.19	14.41	11.32	33.84	32.96
	PT [38]	0.197	6.27	3.33	2.14	1.59	14.52	9.50
	PT + GN [38]	0.191	11.45	7.08	5.08	4.04	19.09	14.52
	NAT-AT [14]	0.177	14.26	9.93	7.11	5.53	21.87	18.72
	NAT-EA [14]	0.146	15.12	10.45	7.99	6.66	22.98	19.43
	PoseVQ-MP [52]	0.146	15.43	10.69	8.26	6.98	-	-
	PoseVQ-DDM [52]	0.116	16.11	11.37	9.22	7.50	-	-
	Stitching G2P (Isolated)	0.593	21.47	8.79	4.25	2.49	23.74	20.32
	Stitching G2P (Continuous)	0.587	23.58	12.31	8.05	5.95	28.85	24.84
	Stitching T2P (Isolated)	0.594	22.78	9.68	5.17	3.12	24.27	21.30
	Stitching T2P (Continuous)	0.572	25.14	13.54	8.98	6.67	29.5	26.49

Using a continuous dictionary we can outperform all models except for the VQ based approaches on BLEU-3 to 4. As the VQ model is learning sub-units of a gloss sequence we suggest this gives it an advantage on higher n-grams, as each token can represent multiple signs.

Segmentation Results To evaluate the stitching segmentation approach we calculate the duration for each gloss in the mDGS dataset test set. We achieve a sign level frame F1-score of 0.6373 a similar score compared to [29] that achieves a top score of 0.63. Demonstrating the effectiveness of stitching for sign segmentation. It is worth noting our approach requires gloss information, but is computationally inexpensive compared to the LSTM used in [29].

5.2 Qualitative Evaluation

Visual Outputs To demonstrate the approach’s effectiveness, we present skeleton and video outputs for two sign languages (BSL and DGS)¹. To enable a fair evaluation, we also share failure cases where incorrect predictions from the initial translation produce repetitive sequences. Furthermore, in the supplementary material, we share visualizations of the produced sequence.

Survey Results The survey presented a comparison to PT, followed by an ablation of different components of the stitching approach. 17% of people surveyed were native Deaf signers, while 34% were L2 signers or language learners.

¹ https://github.com/walsharry/Sign_Stitching_Demos

87.5% preferred our approach compared to the PT, while the rest selected no preference. 100% of people agreed that applying the filter improved the realism compared to no filtering, while resampling was found to be less important, with 37.5% selecting no preference between the resampled and normal sequence. Interestingly, 41.7% preferred using an isolated dictionary, while 20.8% preferred the continuous dictionary and the rest indicated no preference.

6 Conclusion

In this paper, we presented a novel approach to SLP. Previous works have suffered from the problem of regression to the mean and have mainly focused on manual features. Here we have overcome the problem by using a dictionary of expressive examples. The stitching effectively joins the signs together creating a natural continuous sequence and by clustering facial expressions into a vocabulary we can create a sequence that contains both manual and non-manual features. We eliminated unnatural transitions and enhanced the stylistic cohesiveness through the approach. As a result, we present state-of-the-art performance. Finally, the user evaluation agrees with the quantitative results, indicating our approach can produce realistic expressive Sign language.

References

1. Agarap, A.F.: Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375 (2018) [10](#)
2. Bangham, J.A., Cox, S., Elliott, R., Glauert, J.R., Marshall, I., Rankov, S., Wells, M.: Virtual signing: Capture, animation, storage and transmission-an overview of the visicast project. In: IEE Seminar on speech and language processing for disabled and elderly people (Ref. No. 2000/025). pp. 6–1. IET (2000) [3](#)
3. Brentari, D.: A prosodic model of sign language phonology. Mit Press (1998) [2](#)
4. Butterworth, S., et al.: On the theory of filter amplifiers. *Wireless Engineer* **7**(6), 536–541 (1930) [8](#)
5. Camgoz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7784–7793 (2018) [3](#), [11](#)
6. Camgoz, N.C., Koller, O., Hadfield, S., Bowden, R.: 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. pp. 10023–10033 (2020) [3](#), [12](#)
7. Cormier, K., Fenlon, J., Johnston, T., Rentelis, R., Schembri, A., Rowley, K., Adam, R., Woll, B.: From corpus to lexical database to online dictionary: Issues in annotation of the bsl corpus and the development of bsl signbank. In: 5th Workshop on the Representation of Sign Languages: Interactions between Corpus and Lexicon [workshop part of 8th International Conference on Language Resources and Evaluation, Turkey, Istanbul LREC 2012. Paris: ELRA. pp. 7–12 (2012) [11](#)
8. Cox, S., Lincoln, M., Tryggvason, J., Nakisa, M., Wells, M., Tutt, M., Abbott, S.: Tessa, a system to aid communication with deaf people. In: Proceedings of the fifth international ACM conference on Assistive technologies. pp. 205–212 (2002) [3](#)

9. Efthimiou, E., Fotinea, S.E., Hanke, T., Glauert, J., Bowden, R., Braffort, A., Collet, C., Maragos, P., Lefebvre-Albaret, F.: The dicta-sign wiki: Enabling web communication for the deaf. In: Computers Helping People with Special Needs: 13th International Conference, ICCHP 2012, Linz, Austria, July 11-13, 2012, Proceedings, Part II 13. pp. 205–212. Springer (2012) [3](#)
10. ElGhoul, O., Jemni, M.: Websign: A system to make and interpret signs using 3d avatars. In: Proceedings of the Second International Workshop on Sign Language Translation and Avatar Technology (SLTAT), Dundee, UK. vol. 23 (2011) [3](#)
11. Gibet, S., Lefebvre-Albaret, F., Hamon, L., Brun, R., Turki, A.: Interactive editing in french sign language dedicated to virtual signers: Requirements and challenges. *Universal Access in the Information Society* **15**, 525–539 (2016) [4](#)
12. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (2010) [10](#)
13. Heloir, A., Gibet, S.: A qualitative and quantitative characterisation of style in sign language gestures. In: International Gesture Workshop. pp. 122–133. Springer (2007) [2](#)
14. Huang, W., Pan, W., Zhao, Z., Tian, Q.: Towards fast and high-quality sign language production. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 3172–3181 (2021) [2](#), [4](#), [12](#), [14](#)
15. Hwang, E.J., Lee, H., Park, J.C.: Autoregressive sign language production: A gloss-free approach with discrete representations. arXiv preprint arXiv:2309.12179 (2023) [2](#), [4](#)
16. Ivashechkin, M., Mendez, O., Bowden, R.: Improving 3d pose estimation for sign language. In: 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW). pp. 1–5 (2023). <https://doi.org/10.1109/ICASSPW59220.2023.10193629> [4](#), [6](#)
17. Kaiser, L., Bengio, S., Roy, A., Vaswani, A., Parmar, N., Uszkoreit, J., Shazeer, N.: Fast decoding in sequence models using discrete latent variables. In: International Conference on Machine Learning. pp. 2390–2399. PMLR (2018) [4](#)
18. Kennaway, R.: Avatar-independent scripting for real-time gesture animation. arXiv preprint arXiv:1502.02961 (2015) [3](#)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014) [10](#)
20. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) [4](#)
21. Koehn, P., Knowles, R.: Six challenges for neural machine translation. arXiv preprint arXiv:1706.03872 (2017) [4](#)
22. Koller, O., Forster, J., Ney, H.: Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding* **141**, 108–125 (2015). <https://doi.org/10.1016/j.cviu.2015.09.013>, <https://www.sciencedirect.com/science/article/pii/S1077314215002088>, pose & Gesture [3](#)
23. Konrad, R., Hanke, T., Langer, G., Blanck, D., Bleicken, J., Hofmann, I., Jeziorski, O., König, L., König, S., Nishio, R., Regen, A., Salden, U., Wagner, S., Worseck, S., Böse, O., Jahn, E., Schulder, M.: Meine dgs – annotiert. öffentliches korpus der deutschen gebärdensprache, 3. release / my dgs – annotated. public corpus of german sign language, 3rd release (2020). <https://doi.org/10.25592/dgs.corpus-3.0>, <https://doi.org/10.25592/dgs.corpus-3.0> [2](#), [4](#), [11](#)
24. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proceedings of the IEEE* (1998) [3](#)

25. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004) 12
26. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M.G., Lee, J., et al.: Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172 (2019) 4, 6, 11
27. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems* (2013) 11
28. Mori, M., MacDorman, K.F., Kageki, N.: The uncanny valley [from the field]. *IEEE Robotics & automation magazine* 19(2), 98–100 (2012) 4
29. Moryossef, A., Jiang, Z., Müller, M., Ebling, S., Goldberg, Y.: Linguistically motivated sign language segmentation. arXiv preprint arXiv:2310.13960 (2023) 14
30. Oord, A.v.d., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning. arXiv preprint arXiv:1711.00937 (2017) 4
31. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. pp. 311–318 (2002) 12
32. Popović, M.: chrF: character n-gram F-score for automatic MT evaluation. In: Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Hokamp, C., Huck, M., Logacheva, V., Pecina, P. (eds.) *Proceedings of the Tenth Workshop on Statistical Machine Translation*. pp. 392–395. Association for Computational Linguistics, Lisbon, Portugal (Sep 2015). <https://doi.org/10.18653/v1/W15-3049>, <https://aclanthology.org/W15-3049> 12
33. Prillwitz, S., für Deutsche Gebärdensprache und Kommunikation Gehörloser, H.Z.: HamNoSys: Version 2.0; Hamburg notation system for sign languages; an introductory guide. Signum-Verlag (1989) 3
34. Rastgoo, R., Kiani, K., Escalera, S., Sabokrou, M.: Sign language production: A review. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 3451–3461 (2021) 4
35. Reilly, J.S., McIntire, M.L., Seago, H.: Affective prosody in american sign language. *Sign Language Studies* pp. 113–128 (1992) 2
36. Saunders, B., Camgoz, N.C., Bowden, R.: Adversarial training for multi-channel sign language production. arXiv preprint arXiv:2008.12405 (2020) 2, 3, 4, 12
37. Saunders, B., Camgoz, N.C., Bowden, R.: Everybody sign now: Translating spoken language to photo realistic sign language video. arXiv preprint arXiv:2011.09846 (2020) 3, 5, 9
38. Saunders, B., Camgoz, N.C., Bowden, R.: Progressive transformers for end-to-end sign language production. In: *European Conference on Computer Vision* (2020) 3, 4, 9, 11, 12, 13, 14
39. Saunders, B., Camgoz, N.C., Bowden, R.: Mixed signals: Sign language production via a mixture of motion primitives. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 1919–1929 (2021) 2, 3, 4, 12
40. Schembri, A.: British sign language corpus project: Open access archives and the observer’s paradox. In: sign-lang@ LREC 2008. pp. 165–169. European Language Resources Association (ELRA) (2008) 11
41. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* (2014) 10
42. Stoll, S., Camgöz, N.C., Hadfield, S., Bowden, R.: Sign language production using neural machine translation and generative adversarial networks. In: *Proceedings of the 29th British Machine Vision Conference* (2018) 12

43. Stoll, S., Camgoz, N.C., Hadfield, S., Bowden, R.: Text2sign: towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision* **128**(4), 891–908 (2020) [4](#)
44. Tamura, S., Kawasaki, S.: Recognition of sign language motion images. *Pattern Recognition* (1988) [3](#)
45. Vali, M.H., Bäckström, T.: Nsvq: Noise substitution in vector quantization for machine learning. *IEEE Access* **10**, 13598–13610 (2022) [5](#)
46. Van Wyk, D.E.: Virtual human modelling and animation for real-time sign language visualisation (2008) [3](#)
47. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017) [3](#)
48. Ventura, L., Duarte, A., Giró-i Nieto, X.: Can everybody sign now? exploring sign language video generation from 2d poses. *arXiv preprint arXiv:2012.10941* (2020) [5](#), [9](#)
49. Wilbur, R.B.: Stress in a sl: Empirical evidence and linguistic issues. *Language and speech* **42**(2-3), 229–250 (1999) [2](#)
50. Wilbur, R.B.: Effects of varying rate of signing on asl manual signs and nonmanual markers. *Language and speech* **52**(2-3), 245–285 (2009) [1](#)
51. Wolfe, R., McDonald, J.C., Hanke, T., Ebling, S., Van Landuyt, D., Picron, F., Krausneker, V., Efthimiou, E., Fotinea, E., Braffort, A.: Sign language avatars: A question of representation. *Information* **13**(4), 206 (2022) [5](#)
52. Xie, P., Zhang, Q., Li, Z., Tang, H., Du, Y., Hu, X.: Vector quantized diffusion model with codeunet for text-to-sign pose sequences generation. *arXiv preprint arXiv:2208.09141* (2022) [2](#), [4](#), [12](#), [14](#)
53. Zwitterlood, I., Verlinden, M., Ros, J., Van Der Schoot, S., Netherlands, T.: Synthetic signing for the deaf: Esign. In: *Proceedings of the conference and workshop on assistive technologies for vision and hearing impairment (CVHI)* (2004) [3](#)