Non-Autoregressive Sign Language Production with Gaussian Space

Eui Jun Hwang
ejhwang@nlp.kaist.ac.kr
Jung-Ho Kim
jhkim@nlp.kaist.ac.kr
Jong C. Park
park@nlp.kaist.ac.kr

Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, Korea

Abstract

Sign Language Production (SLP) aims to translate spoken language expressions into sign language expressions such as a sequence of sign poses or a sign video. Previous SLP works have used an autoregressive approach to learn the relationship between spoken words and sign poses. However, since the approaches work autoregressively, the decoder unintentionally regresses to the mean and even suffers from error propagation. In this work, we propose Non-Autoregressive Sign Language Production with Gaussian space (NSLP-G), a novel SLP model that uses non-autoregressive decoding to generate sign poses. To avoid direct regression, NSLP-G makes use of two phases. The first phase is to build a pose generator capable of generating various sign poses in a continuous sign pose space. At the second phase, we use a non-autoregressive Transformer to map from the source sentence to the target distribution. To validate the results of our model, we assess the quality of produced sign poses using Fréchet Gesture Distance, Mean Absolute Error of Joint coordination and back-translation evaluation. Experimental results show that NSLP-G outperforms the state-of-the-art model on the RWTH-PHOENIX-Weather 2014T dataset.

1 Introduction

Sign language is the primary language of the Deaf community. Unlike spoken language, sign language conveys the meaning through manual and non-manual elements, which may invoke a communication gap between the Deaf and hearing individuals. To lessen the gap, sign language interpreters have provided translated information interactively, which is still insufficient to meet the high demand for interpretation. As an alternative, sign language researchers have proposed vision-based sign language translation to reduce the dependency on sign language interpreters. Sign Language Production (SLP) translates spoken language expressions into actual sign language expressions such as sign pose sequences, sign animations, or videos. In this work, we aim to generate realistic and continuous sign pose sequences with varying lengths from a sequence of spoken words or sign glosses.

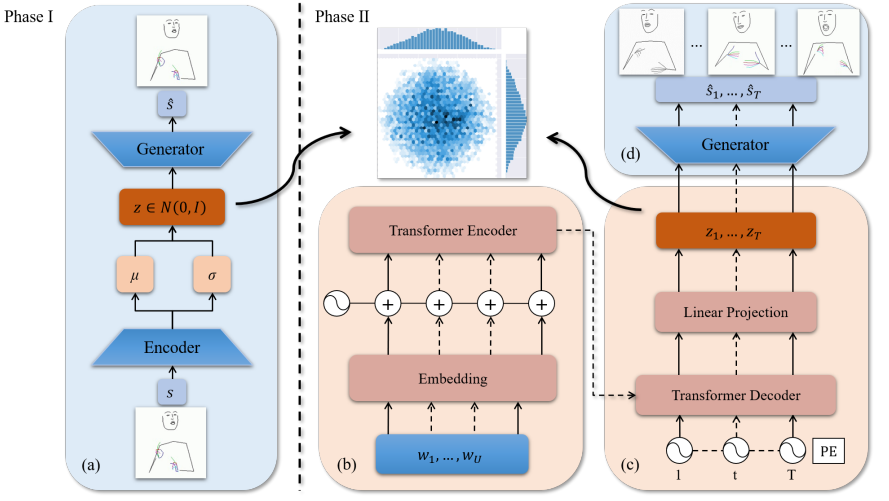

Figure 1: An overview of the proposed SLP model, which consists of two phases to generate sign poses on a given source sentence (spoken language or sign gloss). Phase I: we use VAE to employ as GPG (see (a) and (d)). The encoder outputs μ and σ and then uses a reparameterization trick to sample z following $\mathcal{N}(0,I)$. The decoder reconstructs $\hat{s}$ using z. Phase II: To generate a sequence of z, we employ a Transformer equipped GPG from (a) to produce a sequence of z (see (b) and (c)). To achieve non-autoregressiveness, autoregressive connections are removed from the decoder and only positional encodings (PE) are used as inputs (see (c)).

The main challenge of SLP is to generate a sequence of meaningful and perceptually realistic sign poses. Previous SLP models [26, 34] have taken an autoregressive approach, but typically suffer from the regression to the mean and error propagation problems. To address these limitations, Saunders *et al.* [25, 27] have applied adversarial learning and Gaussian Mixture Density network to their previous work, respectively. Although these efforts have mitigated these problems to some extent, they have not yet been fully resolved.

To address these problems from a more fundamental perspective, we propose a novel SLP model, Non-Autoregressive Sign Language Production with Gaussian space (NSLP-G), which has a completely different approach from the existing SLP models. It makes use of two different phases: building a pose generator and mapping from a source sentence to target sign pose distributions. More specifically, at Phase I, we use Variational Autoencoder (VAE) to perform self-supervised learning on the sign poses. After the learning process, the decoder can generate a sign pose in Gaussian space, which is now employed as Gaussian Pose Generator (GPG). At Phase II, we use non-autoregressive Transformer as Gaussian Seeker (GS) to translate the source sentence to the target sign pose distributions based on GPG. The key novelty of our model is to provide the decoder with positional encoding and output the entire sign pose sequence at once. To the best of our knowledge, there is no such SLP model that produces sign poses in a non-autoregressive manner. We find that a non-autoregressive SLP model produces sign poses accurately given a well-constructed sign pose space. It also mitigates regression to the mean and error propagation in decoding. An overview of our approach is shown in Figure 1.

In previous SLP studies, a back-translation approach [26] evaluates the quality of gen-

erated sign poses. However, the back-translation evaluation could be problematic. When comparing the original and back translation produced by an SLT model, discrepancies may arise due to errors in the actual translation, but they can also be due to errors in the back translation [4]. Here, we use Fréchet Gesture Distance (FGD) and Mean Absolute Error of Joint coordination (MAEJ) proposed for a gesture generation task [32] as evaluation metrics to assess the quality of produced sign poses directly.

The main contributions of this work are as follows:

- We introduce a novel Gaussian space-based SLP model, NSLP-G, which produces sign poses in a non-autoregressive manner.
- To enable a more direct comparison, we introduce new evaluation metrics for SLP models.
- Our proposed model achieves the state-of-the-art performance based on the proposed metrics and the back-translation evaluation.

The rest of our paper is organized as follows. Section 2 reviews the existing literature on SLP and non-autoregressive models. Section 3 introduces the details of NSLP-G. Section 4 provides details of the experimental setup, results, and analysis. Finally, Section 5 concludes the paper and proposes future research directions.

2 Related Work

2.1 Sign Language Production

Previous work on SLP can be divided into four categories: avatar, statistics, neural network, and motion graph based models. The avatar-based models can produce human-like signs, but rely on phrase lookups and pre-defined gesture sequences [11, 16, 23], or require expensive motion capture or pre-recorded phrases [3, 18, 22, 35].

With recent advances in deep learning, Stoll *et al*. [28] propose the first SLP model based on neural machine translation and generative adversarial network. They divide the SLP task into two different phases: translating from natural utterances to sign glosses and mapping the glosses to corresponding sign skeleton poses. Zelinka and Kanis [34] propose a first end-to-end learning method from text to sign poses with a fixed length and ordering. They also introduce an iterative backpropagation method that interpolates missing skeleton joints from the extracted skeleton poses. Along with the interpolation method, Saunders *et al*. [26] propose a transformer-based SLP model that allows a dynamic length of output sign poses, and introduce a counter encoding scheme for learning the start and end of the pose sequences. Although the model performs in a relatively stable manner, the outputs are still under-expressed due to the regression to the mean problem. To address the problem, they adopt an adversarial training method consisting of a generator and a discriminator into the model [25], with slightly better results compared to the previous work.

2.2 Non-Autoregressive Translation

Autoregressive approach has achieved a great success in machine translation [29, 30, 31]. Despite its success, autoregressive approach has two main drawbacks: 1) the autoregressive decoder highly relies on its previous target outputs, resulting in the error propagation from previous predictions [19]. 2) In case of predicting human poses, autoregressive models are prone to converging to a mean pose, which hinders the prediction of realistic poses [20].

Recent studies resolve such problems in a non-autoregressive manner, where Gu *et al*. [13] propose a non-autoregressive decoder, generating a target sequence given a source sentence and a sequence of fertility values. Guo *et al*. [14] enhance the decoder input by directly leveraging a phrase table and adversarial learning. In human motion prediction, Li *et al*. [19] remove autoregressive connections in the decoder, generating each target pose independently given context features from the encoder and positional information.

In our work, we formulate the SLP problem in a non-autoregressive manner, injecting positional information to the decoder in NSLP-G to generate sign poses in parallel. Although Gu *et al*. [13] point out that performance degradation occurs when omitting inputs to the decoder or using only positional encoding, we resolve this problem by equipping a pretrained pose generator to NSLP-G, which guides the model to generate a sequence of sign poses.

3 Non-Autoregressive Sign Language Production with Gaussian Space

3.1 Problem Definition

Suppose that the source sentence of length U is denoted as $W = (w_1, w_2, ..., w_U)$, and that the target sequence of produced sign poses of length T is denoted as $S = (s_1, s_2, ..., s_T)$. The previous SLP works [25, 26, 27, 34] maximize the conditional probability $P(S|W)$.

However, due to the curse of dimensionality [33] and the drastic difference in length between W and S, an autoregressive approach does not work well. To address this problem, we avoid direct regression and instead map words W to sign pose distributions Z, where $Z = (z_1, z_2, .., z_T)$ generates sign pose S using a generator $g(\cdot)$. This can be formulated as:

$$P(Z|W), \quad g(S|Z), \quad z_i \in \mathcal{N}(0, I) \tag{1}$$

To maximize $P(Z|W)$ and $g(S|Z)$, we use Transformer and VAE, respectively. We will cover each method in detail in Subsections 3.2 and 3.3.

3.2 Gaussian Pose Generator with VAE

At Phase I, as shown in Figure 1 (a), we employ VAE, which is widely used for generative tasks [5, 9, 10, 15, 21] to obtain GPG. We train the VAE to generate a sign pose $\hat{s}$ as close as possible to the ground truth sign pose s. It has a simple architecture consisting of sign pose encoder enc_{sp} and decoder dec_{sp} similar to Autoencoder (AE) [2]. The encoder takes a sign pose s and encodes it into latent space z_{sp}. The decoder reconstructs a sign pose from the latent space z_{sp}. The encoder and decoder can be denoted as follows:

$$enc_{sp}(s) = q_{sp}(z_{sp}|s), \quad dec_{sp}(z_{sp}) = p_{sp}(s|z_{sp}), \tag{2}$$

where $q_{sp}(z_{sp}|s)$ and $p_{sp}(s|z_{sp})$ are the posterior distributions for the encoder and decoder, respectively.

VAE uses a reparameterization trick to sample the latent vector z_{sp} from the encoder's output to project the sign pose s into Gaussian space. The reparameterization trick can be formulated as follows:

$$z_{sp} = \mu_{sp} + \sigma_{sp} \odot \varepsilon, \quad where \quad \varepsilon \in \mathcal{N}(0, I), \tag{3}$$

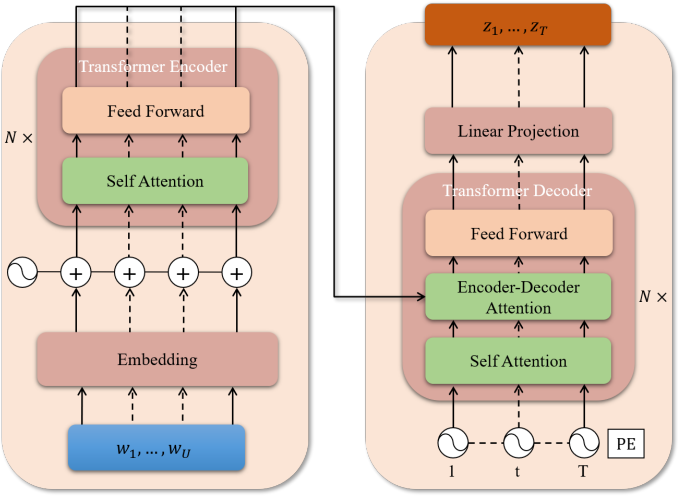

Figure 2: An overview of Transformer based Gaussian Seeker. It consists of a transformer encoder and a non-autoregressive decoder. The encoder takes source sentence $w_1, \ldots, w_U$ with U words and the decoder takes PE with T length as an input and generate a sequence of latent vector Z following a Gaussian distribution.

where μ_{sp} and σ_{sp} are the mean and variance of the sign pose distribution, respectively; ε is an auxiliary independent random variable; and $\odot$ is element-wise multiplication.

The loss function of VAE is formulated as:

$$L_{vae}(s) = -E_{z_{sp}\sim q_{sp}(z_{sp}|s)}[\log p_{sp}(s|z_{sp})] + \beta KL(q_{sp}(z_{sp}|s)||p_{sp}(z_{sp})), \tag{4}$$

where $p(z_{sp}) = \mathcal{N}(0,I)$ is the prior distribution and $KL(\cdot||\cdot)$ is the Kullback-Leibler (KL) divergence. The first term allows the model to encode the sign pose s into the latent space $z_{sp} \in \mathcal{N}(0,I)$ for reconstruction. We use Mean Squared Error (MSE) loss to let the decoder assume Gaussian distribution. The second term pushes posterior distribution $q_{sp}(z_{sp}|s)$ to be close to the prior distribution $p_{sp}(z_{sp})$. We add a variable weight β defined by KL cost annealing [5]. After the learning process, the trained decoder $dec_{sp}(\cdot)$ is defined as GPG (see Figure 1 (d)).

3.3 Gaussian Seeker with Non-Autoregressive Transformer

At Phase II, as shown in Figure 2, we build a Transformer in a non-autoregressive manner and employ it as GS.

Encoder. We use the same Transformer encoder, which consists of a stack of N identical layers with Multi Head Attention (MHA) and Feed Forward layers. The inputs are first embedded into $n-$dimensional space and the positions are added to the embedded representation of each input. MHA performs scaled dot product attention to generate a weighted contextual representation of the given inputs. Scaled dot product attention can be formulated as follows:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{5}$$

Models	DEV		TEST	
	FGD↓	MAEJ* ↓	FGD↓	MAEJ* ↓
Real	$1.59^{\pm 0.30}$	$2.95^{\pm 0.05}$	$1.69^{\pm 0.30}$	$3.19^{\pm 0.01}$
Gloss to Pose (G2P)				
Saunders *et al.* [26]	$2.43^{\pm 0.04}$	$3.76^{\pm 0.01}$	$3.23^{\pm 0.51}$	$3.79^{\pm 0.03}$
Ours	$\mathbf{2.10}^{\pm 0.06}$	$\mathbf{3.45}^{\pm 0.01}$	$\mathbf{2.83}^{\pm 0.06}$	$\mathbf{3.52}^{\pm 0.02}$
Text to Pose (T2P)				
Saunders *et al.* [26]	$2.48^{\pm 0.08}$	$3.89^{\pm 0.04}$	$3.33^{\pm 0.19}$	$4.01^{\pm 0.05}$
Ours (*T2P*)	$2.36^{\pm 0.05}$	$3.64^{\pm 0.02}$	$\mathbf{3.02}^{\pm 0.07}$	$3.76^{\pm 0.02}$
Ours (*T2PG*)	$\mathbf{2.33}^{\pm 0.06}$	$3.63^{\pm 0.03}$	$3.12^{\pm 0.02}$	$3.75^{\pm 0.02}$
Ours (*T2P+finetuning*)	$2.45^{\pm 0.03}$	$3.61^{\pm 0.02}$	$3.23^{\pm 0.01}$	$3.71^{\pm 0.01}$
Ours (*T2PG+finetuning*)	$2.44^{\pm 0.07}$	$\mathbf{3.60}^{\pm 0.02}$	$3.15^{\pm 0.17}$	$\mathbf{3.70}^{\pm 0.01}$

Table 1: A comparison of the performance of our models with state-of-the-art models. The ± shows 95% confidence intervals over tasks and the ↓ indicates that a lower number is better. MAEJ* is scaled 100× for better readability.

where Q, K and V denote query, key and value, respectively. This allows the model to learn the relationship between each input in the sequence and how they relate to each other. Finally, MHA can be formulated as:

$$MHA(Q,K,V) = Concat(head_i,...,head_n)W_O, \tag{6}$$

$$head_i = Attention(QW_Q^i, KW_K^i, VW_V^i), \tag{7}$$

where W_O, W_Q, W_K and W_V are weights related to each input.

Decoder. We remove the autoregressive mask so that the decoder works in a non-autoregressive manner. The $P(Z|W)$ in Equation 1 can be represented as follows:

$$P_{\mathcal{NA}}(Z|W) = \prod_{t=1}^{T} p_{gs}(z_t|w_{1:U}), \tag{8}$$

where Z and W are a target sequence of sign poses and a source sentence, respectively.

These distributions can be computed in parallel at inference time. However, as can be seen from Equation 8, there is no conditional probability to predict the length of the target distribution sequence Z. Our model generates a fixed sequence of sign poses, but at the same time, utilizes a masked MSE loss that enables the model to learn sign poses of variable length. With the loss calculation, our model converges to an idle state when inference is complete.

Specifically, our decoder takes PE as a query and the encoder's output as a key and value. The decoder has the same number of layers as the Encoder does, and each layer contains MHA self-attention (Equation 6), encoder-decoder attention, and feed-forward sub-layers. Finally, the decoder outputs a target sequence Z through a linear projection layer.

4 Experiments

4.1 Experimental Settings

Dataset. We evaluate our proposed models on the publicly available RWTH-PHOENIX-Weather 2014T dataset [6]. The corpus contains 8,257 pairs of German and German Sign

Models	DEV					TEST				
	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE
Real	11.44	14.49	20.12	32.26	33.30	11.30	14.27	19.96	32.37	32.63
Gloss to Pose (G2P)										
Saunders *et al.* [26]	7.18	9.19	13.21	24.33	25.81	6.24	8.17	12.05	22.78	24.55
Ours	**10.28**	**13.04**	**18.13**	**29.83**	**31.38**	**9.39**	**12.12**	**17.30**	**28.98**	**30.38**
Text to Pose (T2P)										
Saunders *et al.* [26]	8.13	10.45	14.88	26.37	27.53	7.60	9.87	14.20	25.27	27.35
Ours (T2P)	10.70	13.46	18.39	29.62	31.37	10.95	**14.07**	19.49	31.23	32.25
Ours (T2PG)	10.98	13.74	18.67	30.40	31.97	10.94	13.84	19.08	30.42	31.92
Ours (T2P+finetuning)	**11.15**	**13.97**	**18.94**	**30.45**	32.09	**11.07**	14.06	**19.57**	**31.68**	**32.63**
Ours (T2PG+finetuning)	11.00	13.81	18.84	30.40	**32.13**	10.88	13.79	18.94	30.06	31.32

Table 2: A comparison of the back-translation evaluation [26] of our model with state-of-the-art models. Note that due to differences in implementation (i.e., lifted sign pose data, number of epochs used) the metrics for the baselines differ from those reported in their paper. The performance improvement with our NSLP-G shows a clear gap from the baselines.

Models	DEV		TEST	
	FGD↓	MAEJ* ↓	FGD↓	MAEJ* ↓
Real	$1.59^{\pm 0.30}$	$2.95^{\pm 0.05}$	$1.69^{\pm 0.30}$	$3.19^{\pm 0.01}$
G2P, latent size = 16	$2.15^{\pm 0.05}$	$3.62^{\pm 0.00}$	$3.10^{\pm 0.15}$	$3.72^{\pm 0.02}$
G2P, latent size = 32	$\mathbf{2.10}^{\pm 0.06}$	$\mathbf{3.45}^{\pm 0.01}$	$\mathbf{2.83}^{\pm 0.06}$	$\mathbf{3.52}^{\pm 0.02}$
G2P, latent size = 64	$2.11^{\pm 0.03}$	$3.62^{\pm 0.01}$	$2.84^{\pm 0.09}$	$3.71^{\pm 0.02}$
G2P, latent size = 128	$2.11^{\pm 0.01}$	$3.61^{\pm 0.02}$	$3.30^{\pm 0.60}$	$3.79^{\pm 0.00}$

Table 3: Effect of Gaussian latent size

Language (DGS) videos with word-level annotations, collected from weather forecasts of PHOENIX TV station. Specifically, the corpus covers 2,887 different German words and 1,066 different DGS glosses. Note that our models are trained to generate skeletal joint coordinate values of sign poses. We use OpenPose [8] to extract 2D manual features of each video and then lift them into 3D using a skeletal correction model [34]. For non-manual features, we also use OpenPose to extract 70 face landmarks represented in 2D coordinates. The landmarks are then scaled to a consistent size and centered to the neck joint.

Evaluation metrics. The back-translation evaluation is limited in measuring the performance of the generated sign poses. This is because it relies heavily on the performance of the SLT model [7], and the translation performance is not yet stable enough to warrant the SLP models (BLEU-4 score of 9.94). To the best of our knowledge, the back-translation evaluation model is not publicly available, making it harder for an exact comparison.

For the sake of reproducibility and fair comparison, we use **Fréchet Gesture Distance (FGD)** and **Mean Absolute Error of Joint coordination (MAEJ)** [32] as evaluation metrics to assess the quality of produced sign poses. FGD between the Gaussian mean and covariance of the latent features of real sign poses S and those of the latent features of the produced sign poses $\hat{S}$ can be represented as follows:

$$FGD(S,\hat{S}) = \left\|\mu_r - \mu_g\right\|^2 + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{\frac{1}{2}}), \tag{9}$$

where μ_r and Σ_r are the first and second moments of the latent feature distribution Z_r of real sign poses S, respectively, and μ_g and Σ_g are the first and second moments of the latent feature distribution Z_g of produced sign poses $\hat{S}$, respectively.

FGD measures the diversity of produced sign poses, whereas MAEJ independently measures the distance between the produced sign pose and real sign pose (no aspect of temporal distance). As the metrics require features from sign poses, we use Transformer AE as a

Models	DEV		TEST	
	FGD↓	MAEJ* ↓	FGD↓	MAEJ* ↓
Real	$1.59^{\pm 0.30}$	$2.95^{\pm 0.03}$	$1.69^{\pm 0.30}$	$3.19^{\pm 0.03}$
G2P (w/ GPG)	$\mathbf{2.10}^{\pm 0.06}$	$\mathbf{3.45}^{\pm 0.01}$	$\mathbf{2.83}^{\pm 0.06}$	$\mathbf{3.52}^{\pm 0.02}$
G2P (w/o GPG)	$18.90^{\pm 0.05}$	$22.71^{\pm 0.02}$	$21.53^{\pm 0.12}$	$25.67^{\pm 0.01}$

Table 4: Effect of GPG

Models	DEV		TEST	
	FGD↓	MAEJ* ↓	FGD↓	MAEJ* ↓
Real	$1.59^{\pm 0.30}$	$2.95^{\pm 0.03}$	$1.69^{\pm 0.30}$	$3.19^{\pm 0.03}$
G2P (non-autoregressive)	$\mathbf{2.10}^{\pm 0.06}$	$\mathbf{3.45}^{\pm 0.01}$	$\mathbf{2.83}^{\pm 0.06}$	$\mathbf{3.52}^{\pm 0.02}$
G2P (autoregressive)	$8.05^{\pm 1.70}$	$4.39^{\pm 0.23}$	$9.10^{\pm 2.10}$	$4.60^{\pm 0.32}$

Table 5: Non-Autoregressive vs. Autoregressive settings

feature extractor. Further details are provided in the supplementary material.

Model settings. For all experiments except the ablation study in Subsection 4.2, GPG has 2 linear layers with ReLU [1]. For GS, we set the embedding dimension to 512, the number of layers to 4, the dropout rate to 0.1, the number of heads to 4, the dimension of the feedforward network to 1024, and the gloss supervision rate to 10^{-5}. All parts of our network are trained with Xavier initialization [12] and Adam optimization [17], with learning rate of 10^{-4}. Our models are implemented using PyTorch [24].

Model types. We employ three different types of models: *Gloss to Pose (G2P)*, *Text to Pose (T2P)*, and *Text to Pose with Gloss supervision (T2PG)*. *G2P* translates a sequence of DGS sign glosses into a sequence of DGS sign poses, whereas *T2P* translates a German sentence. *T2PG* is similar to *T2P* but has an additional decoder that guides its representation to gloss representation.

4.2 Quantitative Results

Comparison with state-of-the-art models. We assess our models on two tasks: *Gloss to Pose (G2P)* and *Text to Pose (T2P)*. For the *G2P* task, we use our *G2P* model, and for the *T2P* task, we use our *T2P* and *T2PG* models. As Progressive Transformer (PT) [26] is the only publicly available SLP model that is trained on the RWTH-PHOENIX-Weather 2014T dataset, we compare the three types of our models with *G2P* and *T2P* models of PT. Note that all models generate the same structure of the sign pose including face, upper body, and hands. We follow the original implementation of PT models and choose the Gaussian noise augmented model as a representative model of PT. These models are trained for 300 epochs. Overall, more epochs may improve performance, but we stop training to keep computational costs low.

In order to measure FGD and MAEJ, we extract features from all the produced sign poses and real (ground truth) sign poses. As shown in Table 1, all types of our models, NSLP-G, outperform the baselines which works in an autoregressive manner. The main advantage of NSLP-G is that it uses a well-constructed Gaussian space to produce sign poses in parallel. We further report the performance of the models using the back-translation evaluation introduced in [26]. As shown in Table 2, our models also outperform the autoregressive model. Since the back-translation model trained for SLP is not publicly available, we train the model based on the same hyper-parameters provided by [7]. For a fair comparison, we use the same

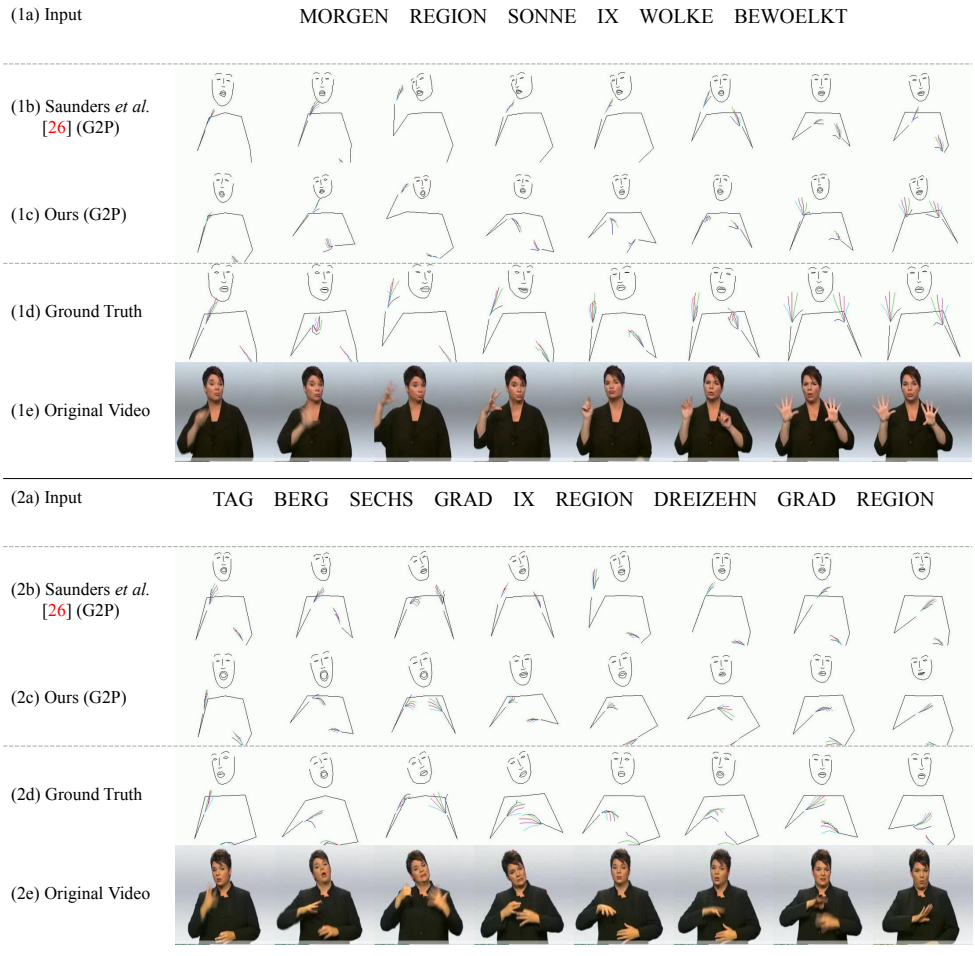

Figure 3: Qualitative results of G2P. We compare our *G2P* model with the state-of-the-art *G2P* model [26]. We uniformly selected 8 frames for each case.

lifted sign poses and the number of epochs for the baselines and our models. Therefore, we have re-assessed the back-translation scores of the baselines, not using the reported scores in [26].

In the experiments, we find that GPG successfully guides GS, producing more expressive and articulate sign pose sequences than baselines. Moreover, the fine-tuned model shows better performance in both evaluation settings. This is because GPG, which assumes a Gaussian distribution, forms better distribution while fine-tuning and consequently gives better results.

Ablation study. To understand NSLP-G in detail, we conduct an ablation study to verify several architectural choices of the model in a controlled setting. Note that the G2P model is used only in this study. As shown in Table 3, we assess our model with different Gaussian latent sizes. The best performance is achieved with the latent size of 32, which is empirically set to an optimal size to contain the necessary information to guide GS.

We also examine the effect of GPG on our model. As shown in Table 4, our G2P model

with GPG produces more realistic sign poses than that without GPG in terms of both FGD and MAEJ. We demonstrate that GPG effectively guides GS. To explore the effectiveness of the non-autoregressive approach in NSLP-G, we also compare the performance between the non-autogressive and autogressive settings as shown in Table 5. Non-autoregressive setting results in better performance than autoregressive setting in our model. Further ablations are available in the supplementary material.

4.3 Qualitative Results

Figure 3 shows two cases of the *G2P* translation. We compare the sign pose sequences produced by our model with those done by Saunders *et al*. [26]. In the first case, our model successfully translates the gloss sequence (see 1a) to the sign pose sequence (see 1c) except fourth and fifth sign poses, which means that the non-autoregressive decoder of our model successfully produces the next sign pose without propagating errors from the previously produced sign poses.

The second case shows another effect of our non-autoregressive decoding. Our model generates more dynamic and accurate sign poses (see 2c), especially facial expressions, than the state-of-the-art model (see 2b). This demonstrates that our model resolves the regression to the mean problem mentioned in the most recent SLP work by Saunders *et al*. [27]. Overall, the results show that our model exploits the nature of non-autoregressive decoding and thus produces more realistic sign poses.

5 Conclusions

In this paper, we propose NSLP-G, the first Gaussian space-based SLP model, to generate sign poses in a non-autoregressive manner. To achieve this, we separate the learning process into building the sign pose generator and mapping the source sentence to the sign pose distributions. NSLP-G then takes source sentences and temporal information to generate sign poses in parallel. Moreover, we introduce FGD and MAEJ to evaluate the produced sign poses more quantitatively. The extensive experiments demonstrate the superiority of the proposed model over the existing state-of-the-art SLP model.

As future work, we would like to expand the pose generator to produce spatio-temporal sign poses using Transformer VAE. Furthermore, we plan to develop a new metric to evaluate the semantic meaning of the produced sign poses more precisely.

Acknowledgement

The authors would like to thank the anonymous reviewers for their through and valuable comments, as well as Jae Young Lee and Dongyeun Lee for their assistance in preprocessing the data. This work was supported in part by the Institute for Information and communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. 2018-0-00582, Prediction and augmentation of the credibility distribution via linguistic analysis and automated evidence document collection) and in part by the Technology Innovation Program (20014406, Development of interactive sign language interpretation service based on artificial intelligence for the hearing impaired) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea).

References

[1] Abien Fred Agarap. Deep Learning using Rectified Linear Units (ReLU). *arXiv preprint arXiv:1803.08375*, 2018.

[2] Pierre Baldi. Autoencoders, Unsupervised Learning, and Deep Architectures. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 37–49, 2012.

[3] J. Andrew Bangham, S.J. Cox, Ralph Elliott, John R.W. Glauert, Ian Marshall, Sanja Rankov, and Mark Wells. Virtual Signing: Capture, Animation, Storage and Transmission - An Overview of the Visicast Project, 2000.

[4] Dorothée Behr. Assessing the use of Back Translation: The Shortcomings of Back Translation as a Quality Testing Method. *International Journal of Social Research Methodology*, 20(6):573–584, 2017.

[5] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating Sentences from a Continuous Space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany, 2016.

[6] Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural Sign Language Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7784–7793, 2018.

[7] Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10023–10033, 2020.

[8] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-person 2D Pose Estimation using Part Affinity Fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019.

[9] Xingyi Cheng, Weidi Xu, Taifeng Wang, Wei Chu, Weipeng Huang, Kunlong Chen, and Junfeng Hu. Variational Semi-Supervised Aspect-Term Sentiment Analysis via Transformer. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 961–969, Hong Kong, China, 2019.

[10] Le Fang, Tao Zeng, Chaochun Liu, Liefeng Bo, Wen Dong, and Changyou Chen. Transformer-based Conditional Variational Autoencoder for Controllable Story Generation. *arXiv preprint arXiv:2101.00828*, 2021.

[11] J.R.W. Glauert, Ralph Elliott, S.J. Cox, Judy Tryggvason, and Mary Sheard. Vanessa–A System for Communication between Deaf and Hearing People. *Technology and Disability*, 18(4):207–216, 2006.

[12] Xavier Glorot and Yoshua Bengio. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[13] Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. Non-autoregressive Neural Machine Translation. *arXiv preprint arXiv:1711.02281*, 2017.

[14] Junliang Guo, Xu Tan, Di He, Tao Qin, Linli Xu, and Tie-Yan Liu. Non-autoregressive neural machine translation with enhanced decoder input. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3723–3730, 2019.

[15] Junyan Jiang, Gus G Xia, Dave B Carlton, Chris N Anderson, and Ryan H Miyakawa. Transformer VAE: A Hierarchical Model for Structure-aware and Interpretable Music Representation Learning. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 516–520. IEEE, 2020.

[16] Kostas Karpouzis, George Caridakis, S.-E. Fotinea, and Eleni Efthimiou. Educational Resources and Implementation of a Greek Sign Language Synthesis Architecture. *Computers & Education*, 49(1):54–74, 2007.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In Yoshua Bengio and Yann LeCun, editors, *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[18] Michael Kipp, Alexis Heloir, and Quan Nguyen. Sign Language Avatars: Animation and Comprehensibility. In *International Workshop on Intelligent Virtual Agents*, pages 113–126. Springer, 2011.

[19] Bin Li, Jian Tian, Zhongfei Zhang, Hailin Feng, and Xi Li. Multitask Non-Autoregressive Model for Human Motion Prediction. *IEEE Transactions on Image Processing*, 30:2562–2574, 2021. doi: 10.1109/TIP.2020.3038362.

[20] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional Sequence to Sequence Model for Human Dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5226–5234, 2018.

[21] Zhaojiang Lin, Genta Indra Winata, Peng Xu, Zihan Liu, and Pascale Fung. Variational Transformers for Diverse Response Generation. *arXiv preprint arXiv:2003.12738*, 2020.

[22] Pengfei Lu and Matt Huenerfauth. Data-driven Synthesis of Spatially Inflected Verbs for American Sign Language Animation. *ACM Transactions on Accessible Computing (TACCESS)*, 4(1):1–29, 2011.

[23] John McDonald, Rosalee Wolfe, Jerry Schnepp, Julie Hochgesang, Diana Gorman Jamrozik, Marie Stumbo, Larwan Berke, Melissa Bialek, and Farah Thomas. An Automated Technique for Real-time Production of Lifelike Animations of American Sign Language. *Universal Access in the Information Society*, 15(4):551–566, 2016.

[24] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic Differentiation in PyTorch, 2017.

[25] Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Adversarial Training for Multi-Channel Sign Language Production. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2020.

[26] Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Progressive Transformers for End-to-End Sign Language Production. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[27] Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Continuous 3D Multi-Channel Sign Language Production via Progressive Transformers and Mixture Density Networks. *International Journal of Computer Vision*, pages 1–23, 2021.

[28] Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. Sign Language Production using Neural Machine Translation and Generative Adversarial Networks. In *Proceedings of the 29th British Machine Vision Conference (BMVC)*. British Machine Vision Association, 2018.

[29] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to Sequence Learning with Neural Networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, page 3104–3112, 2014.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6000–6010, 2017.

[31] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[32] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity. *ACM Transactions on Graphics (TOG)*, 39(6):1–16, 2020.

[33] Fajrian Yunus, Chloé Clavel, and Catherine Pelachaud. Sequence-to-Sequence Predictive Model: From Prosody To Communicative Gestures. *arXiv preprint arXiv:2008.07643*, 2020.

[34] Jan Zelinka and Jakub Kanis. Neural Sign Language Synthesis: Words are our Glosses. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3395–3403, 2020.

[35] Inge Zwitserlood, Margriet Verlinden, Johan Ros, and Sanny Van Der Schoot. Synthetic Signing for the Deaf: eSIGN. In *Proceedings of the Conference and Workshop on Assistive Technologies for Vision and Hearing Impairment*, 2004.