Supplementary Material for NSLP-G: 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

1 Introduction

This document contains the following:

- Details of the feature extractor used in our experiments
- An additional ablation study of our Sign Language Production (SLP) models

2 Feature Extractor for SLP Evaluation

This section provides details of the feature extractor used to evaluate the SLP models. As shown in Figure 1, we use Transformer based Autoencoder, which is the same architecture as Gaussian Seeker (GS) in the main paper. A temporal pooling layer is added at the end of the encoder to obtain one latent gesture feature space. The encoder part is used as the feature extractor after the learning process. To maximize its performance, we use a fixed length of sign poses and KL weight, set to 32 and 10^{-5}, respectively. The rest of the configurations is the same as Gaussian Seeker (GS) in the main paper.

3 Additional Ablation Study

In this section, we provide an additional ablation study. We find that batch size has a significant impact on model performance. As shown in Table 1, we obtain the best performance when the batch size is set to 64, which is used for all our experiments. Next, we experiment with the number of layers in both the encoder and decoder. Table 2 summarizes the results. Our *Gloss to Pose (G2P)* models with four layers have achieved the best performance, while those with more or fewer layers degrade their performance. Lastly, we study how much gloss supervision rate affects our *Text to Pose with Gloss supervision (T2PG)* models. As can be

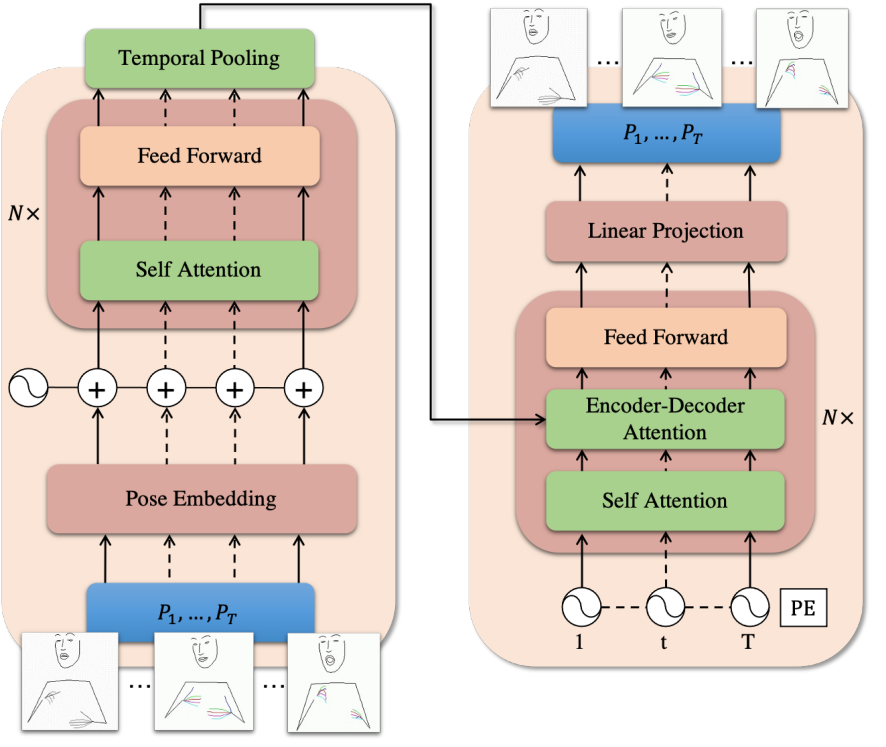

Figure 1: An overview of feature extractor. We use Transformer based autoencoder which has the same architecture as GS in the main paper except for a temporal pooling layer. After learning process, the encoder is used to extract the latent gesture feature space from a given sequence of sign poses.

seen in Table 3, we have found that the gloss supervision rate has a negligible effect on the performance of our *T2PG* models.

Methods	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, batch size = 16	$4.18^{\pm 0.01}$	$3.90^{\pm 0.00}$	$4.73^{\pm 0.03}$	$3.94^{\pm 0.02}$
G2P, batch size = 32	$2.53^{\pm 0.00}$	$3.70^{\pm 0.01}$	$3.07^{\pm 0.03}$	$3.75^{\pm 0.02}$
G2P, batch size = 64	$\mathbf{2.10}^{\pm 0.06}$	$\mathbf{3.45}^{\pm 0.01}$	$\mathbf{2.83}^{\pm 0.06}$	$\mathbf{3.52}^{\pm 0.02}$
G2P, batch size = 128	$2.42^{\pm 0.13}$	$3.75^{\pm 0.02}$	$4.16^{\pm 0.32}$	$3.92^{\pm 0.00}$

Table 1: Effect of the batch size

Methods	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, 2 layers	$2.15^{\pm 0.15}$	$3.52^{\pm 0.02}$	$3.10^{\pm 0.16}$	$3.62^{\pm 0.03}$
G2P, 4 layers	$\mathbf{2.10}^{\pm 0.06}$	$\mathbf{3.45}^{\pm 0.01}$	$\mathbf{2.83}^{\pm 0.06}$	$\mathbf{3.52}^{\pm 0.02}$
G2P, 6 layers	$2.33^{\pm 0.03}$	$3.62^{\pm 0.15}$	$2.84^{\pm 0.08}$	$3.71^{\pm 0.15}$
G2P, 8 layers	$2.29^{\pm 0.13}$	$3.63^{\pm 0.02}$	$2.99^{\pm 0.05}$	$3.73^{\pm 0.00}$

Table 2: Effect of the number of layers

Methods	DEV		TEST	
	FGD↓	MAEJ* ↓	FGD↓	MAEJ* ↓
Real	$1.59^{\pm 0.30}$	$2.95^{\pm 0.05}$	$1.69^{\pm 0.30}$	$3.19^{\pm 0.01}$
T2PG, gsr=10^{-3}	$2.29^{\pm 0.02}$	$3.61^{\pm 0.03}$	$3.38^{\pm 0.22}$	$3.76^{\pm 0.03}$
T2PG, gsr=10^{-4}	$\mathbf{2.28}^{\pm 0.06}$	$\mathbf{3.51}^{\pm 0.04}$	$3.33^{\pm 0.12}$	$3.77^{\pm 0.01}$
T2PG, gsr=10^{-5}	$2.33^{\pm 0.06}$	$3.63^{\pm 0.03}$	$\mathbf{3.12}^{\pm 0.02}$	$\mathbf{3.75}^{\pm 0.02}$
T2PG, gsr=10^{-6}	$2.29^{\pm 0.03}$	$3.53^{\pm 0.01}$	$3.45^{\pm 0.20}$	$3.78^{\pm 0.06}$

Table 3: Effect of the gloss supervision rate. gsr denotes the gloss supervision rate.