Appendices

A Details of Sign Language Representation

We calculate the 6D representation of each joint point as follows:

$$r_k^{\rightarrow} = gGS(\exp(\vec{\omega}_k)) \tag{6}$$

$$\exp(\vec{\omega}_k) = \mathcal{I} + \widehat{\bar{\omega}}_k \sin(\|\vec{\omega}_k\|) + \widehat{\bar{\omega}}_k^2 \cos(\|\vec{\omega}_k\|) \tag{7}$$

We first convert the axis-angle representation of each joint point into a rotation matrix using the Rodrigues formula, where $\bar{\omega} = \frac{\vec{\omega}}{\|\vec{\omega}\|}$ denote the unit norm axis of rotation, $\widehat{\bar{\omega}}$ is the skew symmetric matrix of the 3 -vector $\bar{\omega}$ and $\mathcal{I}$ is the 3×3 identity matrix. Then we use the mapping function gGS defined in (Zhou et al., 2019), which drop the last column of the rotation matrix and flattens it, to convert the rotation matrix into a 6D representation.

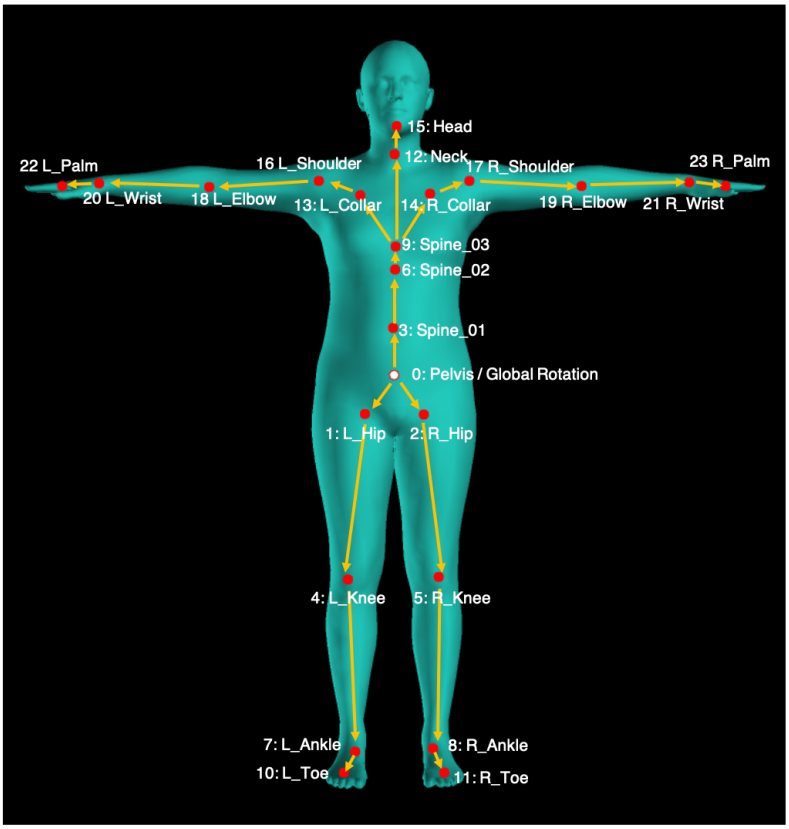

Figure 5: The body skeleton used by the frankmocap model.

The body skeleton used by the frankmocap model is shown in Figure 5. In NaturalSigner, we do not use joints 0, 1, 2, 4, 5, 7, 8, 10, 11, 22 and 23, because joints 22 and 23 do not exist in the SMPL-X model, joint 0 represents the overall rotation, and other joints belong to the invisible part of the lower body. For the hands, we use all the joints in the SMPL-X model.

B Implementation Details

B.1 Model Configurations

For the gloss encoder and text encoder in the mixed semantic encoder, we use vanilla Transformer encoders. For text and gloss, we use space-separated words as input. For the gloss to text attention

module, we use the same setting as the multi-head attention in the gloss encoder. We list the hyper-parameters of NaturalSigner in Table 7. For text to gloss translation, we use the Transformer tiny model architecture in the fairseq toolkit, we list the hyper-parameters of the Translator model in Table 8.

Table 7: The hyper-parameters of NaturalSigner.

Hyper-parameter		NaturalSigner
Sign Language Prompt Encoder	Encoder Layers	3
	Hidden Size	512
	Heads	4
	Conv1D Kernel	3
	Conv1D Filter Size	1024
	Prompt Embeding Size	512
	Dropout Rate	0.1
NaturalSigner	Text/Gloss Embedding	512
	Text/Gloss Encoder Layers	4
	Encoder Hidden Size	512
	Encoder Heads	4
	Encoder FeedForwad Size	1024
	Denoiser Layers	8
	Denoiser Hidden Size	512
	Denoiser Heads	4
	Denoiser Conv1D Kernel	3
	Denoiser Conv1D Filter Size	1024
	Dropout Rate	0.1

Table 8: The hyper-parameters of text-to-gloss translator.

Hyper-parameter		Translator
Translator	Encoder/Decoder Layers	2
	Hidden Size	64
	Heads	2
	FeedForwad Size	64
	Dropout Rate	0.3

B.2 Training Details

For the text-to-gloss translator, we use the Adam optimizer for training, the initial learning rate is 5×10^{-4} (β_1=0.9, β_2=0.998, $\epsilon = 10^{-8}$), label smoothing is 0.1, max tokens is 16384, max-update is 40000, warmup-updates is 4000, warmup-init-lr is 1×10^{-7}, dropout is 0.3, and the training takes 0.5 hours. For NaturalSigner, we use the Adam optimizer for training, the initial learning rate is 1×10^{-4}. We used a batch size of 64 and applied noising steps T=1000 using a cosine noise schedule. A total of 600000 steps were updated during training, which lasted for 48 hours. In subsection 3.4, the weights of different losses are $\lambda_1 = 1$, $\lambda_2 = 1$ and $\lambda_3 = 1$ respectively.

C Evaluation Metrics

Following the approach of (Saunders et al., 2020b), we train a back translation model and calculate ROUGH-L, BLEU-1, BLEU-2, BLEU-3, BLEU-4 metrics between back-translation text and ground-truth input text. Inspired by (Lee et al., 2019), we train a feature extraction model under a contrastive learning assumption and calculate FID, Multimodal Distance and Diversity metrics to evaluate the generation quality and diversity of SLG models.

For the non-deterministic models such as NaturalSigner, we repeat experiment 10 times and calculate average score and 95% confidence interval.

We will show details of implementation in this section.

Back Translation. Similar to (Saunders et al., 2020b), we use the open-source sign language translation model (Camgoz et al., 2020) and full data to train a back translation model for evaluating the semantic accuracy of generated sign language pose sequences. We keep the model hyperparameters exactly the same as those in the official implementation configuration.

Text, Gloss and Sign Language Feature Extractor. We train a contrastive learning-based feature extractor for both text/gloss and sign language modalities. For the Phoenix2014T dataset, we extract text and sign language features, while for the Phoenix2014 dataset, gloss and sign language features are extracted since there is no text data available in this dataset. We utilize a simple transformer encoder as feature encoder, then we have:

$$F_m = Linear_{m,2}(Encoder_m(Linear_{m,1}(M))) \tag{8}$$

$$F_s = Linear_{s,2}(Encoder_s(Linear_{s,1}(Embedding_s(S)))) \tag{9}$$

Where S and M represent text/gloss input and sign language pose sequences input respectively. We set $embedding_{dim} = model_{dim} = 512$, $num_{layers} = 4$ and $num_{attn\_head} = 4$ for the both transformer encoders. We train feature extractor using full data samples and a contrastive learning loss(Radford et al., 2021):

$$\mathcal{L}_{\text{contras}} = \text{ContrastiveLoss}(F_m, F_s) \tag{10}$$

Frechet Inception Distance. (FID) We extract features of ground-truth sign language and model-generated sign language. Then we can calculate Frechet Inception Distance (FID)(Heusel et al., 2017) between the distributions of ground-truth sign language and model-generated sign language of all samples in testing set:

$$Score_{\text{FID}} = FID\left(\{F_{m,gt}\}, \{F_{m,gen}\}\right) \tag{11}$$

Where $F_{m,gt}$ and $F_{m,gen}$ represent ground-truth sign language features and model-generated ones. FID metric is widely used to evaluate quality of generation in the feature space.

Multimodal Distance. Since we minimize the distance between sign language feature and text/gloss feature when training feature extractors, the model-generated sign language features should also be close to ground-truth text features. We calculate Multimodal Distance with average Euclidean distance over all samples in testing set:

$$Score_{mm-dist} = \frac{1}{N}\sum_{i=1}^{N}\left\|F_{s,gt}^{i}, F_{m,gen}^{i}\right\|_2 \tag{12}$$

Where N is the size of testing set. This metric can measure the semantic similarity of cross-modal generation.

Diversity. Diversity measures the variance of the generated sign language across all text/gloss. We randomly sample two subsets with the same size S_d from testing set $F_{m,gen}^{i_1}, ..., F_{m,gen}^{i_{S_d}}$ and $F_{m,gen}^{j_1}, ..., F_{m,gen}^{j_{S_d}}$ in random permutation order, combine them to get S_d pairs, then calculate average Euclidean distance of :

$$Score_{divers} = \frac{1}{S_d}\sum_{k=1}^{S_d}\left\|F_{m,gen}^{i_k}, F_{m,gen}^{j_k}\right\|_2 \tag{13}$$

Where $F_{m,gen}^{i_k}$ and $F_{m,gen}^{j_k}$ are two random subsets in random order of one testing result set, and $F_{m,gen}^{i_k}$ represent model-generated sign language feature of i_k-th sample. We sample $S_d = 300$ examples each experiment. We repeat this experiment 10 times and report average score and 95% confidence interval. This metric reflects the overall diversity of a set of sign language data, and the generated sign language are considered better if they are close to the ground truth sign language in this metric.

D ADDITIONAL EXPERIMENTS

Table 9: Comparison of the experimental results of the model under the gloss to pose and text to pose settings on the Phoenix2014T dataset. $\rightarrow$ means results are better if the metric is closer to the real distribution and ± indicates the 95% confidence interval.The range of the interval is multiplied by 10^2 for ease of data display. G2P means gloss to pose setting and T2P means text to pose setting.

(a) Experimental results on validation set.

Method	FID↓	Multimodal Dist ↓	Diversity →
Real	$0.000^{\pm.00}$	$0.651^{\pm.00}$	$0.798^{\pm.23}$
PT-G2P (Saunders et al., 2020b)	$1.963^{\pm.00}$	$1.418^{\pm.00}$	$0.065^{\pm.07}$
PT-T2P (Saunders et al., 2020b)	$1.984^{\pm.00}$	$1.423^{\pm.00}$	$0.061^{\pm.09}$
NaturalSigner-G2P	$0.161^{\pm.43}$	$1.021^{\pm.26}$	$0.755^{\pm.08}$
NaturalSigner-T2P	$\mathbf{0.136^{\pm.10}}$	$\mathbf{0.980^{\pm.19}}$	$\mathbf{0.7637^{\pm.08}}$

(b) Experimental results on test set.

Method	FID↓	Multimodal Dist ↓	Diversity →
Real	$0.000^{\pm.00}$	$0.645^{\pm.00}$	$0.646^{\pm.17}$
PT-G2P (Saunders et al., 2020b)	$1.929^{\pm.00}$	$1.406^{\pm.00}$	$0.053^{\pm.06}$
PT-T2P (Saunders et al., 2020b)	$1.948^{\pm.00}$	$1.411^{\pm.00}$	$0.0496^{\pm.09}$
NaturalSigner-G2P	$0.151^{\pm.23}$	$1.023^{\pm.23}$	$0.607^{\pm.07}$
NaturalSigner-T2P	$\mathbf{0.125^{\pm.13}}$	$\mathbf{0.980^{\pm.20}}$	$\mathbf{0.618^{\pm.06}}$

D.1 COMPARISON OF GLOSS TO POSE AND TEXT TO POSE SETTINGS

We compared the experimental results of the model on the Phoenix2014T dataset under the gloss to pose and text to pose settings, as shown in Table 9. It can be seen that for NaturalSigner, the model under the text to pose setting is better than the model under the gloss to pose setting in terms of FID, Multimodal Dist and Divsersity indicators. For PT, the model under the text to pose setting is worse than the model under the gloss to pose setting in all indicators. This indicates that using translated gloss and ground truth text as conditions can generate more high-quality and correct sign language than using ground truth gloss as conditions. In addition, this also proves the effectiveness of the mixed semantic encoder designed in NaturalSigner.