Supplementary Materials for "G2P-DDM: Generating Sign Pose Sequence from Gloss Sequence with Discrete Diffusion Model"

A. Structured Prediction Layer for Sign Skeleton

In this part, we illustrate the hierarchy chains of the pose in Fig. 1 and the hand in Fig. 2. The Structured Prediction Layer (SPL) models the structure of the skeleton and hence the spatial dependencies between joints.

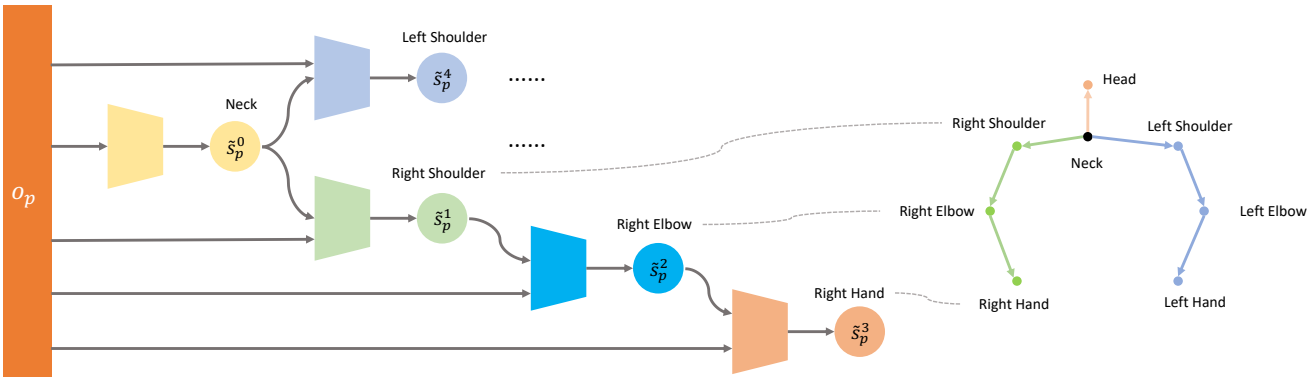

Figure 1: SPL for pose joints. Given the pose feature o_p, joint prediction $\tilde{s}_p^{(k)}$ are made hierarchically by following the spatial chain defined by the underlying skeleton.

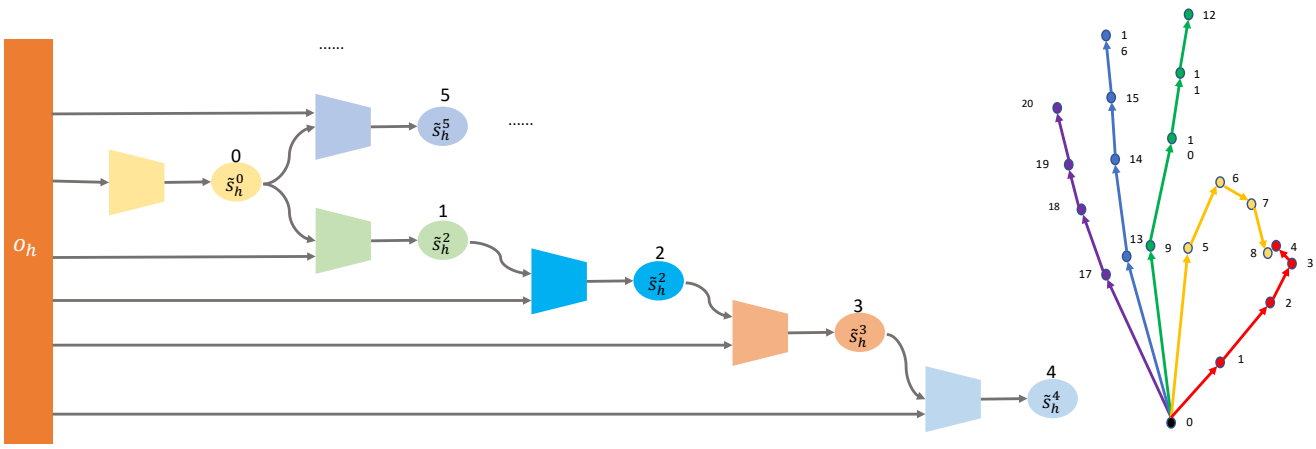

Figure 2: SPL for hand joints. Given the left hand feature or right hand feature o_h, joint prediction $\tilde{s}_h^{(k)}$ are made hierarchically by following the spatial chain defined by the underlying skeleton.

B. Detail of Model architecture

In our experiments for conditional sign pose sequence generation, the input for Pose-VQVAE model is the sign skeleton sequences with 50 joints every frame, where 8 joints for pose, 21 joints for left hand and 21 joints for right hand. Every joint is represented by x, y, z coordinate values.

B.1. Pose-VQVAE

The hyperparameters settings and the details of Pose-VQVAE is shown in Table 1.

Encoder and Decoder	
Input size	$T \times 50 \times 3$
Units of Linear Layer	256
Latent size	$T \times 3 \times 256$
Spatial Transformer layers	3
Temporal Transformer layers	3
Codebook	
Embedding size	256
β (commitment loss coefficient)	0.25
Codebook size	3 x 1024
Others	
Batch size per GPU	6
Optimizer	AdamW
Learning rate	3e-4

Table 1: Hyperparameters of Pose-VQVAE.

B.2. PoseVQ-Diffusion

The hyperparameters settings and the details of Pose-VQDiffusion is shown in Table 2.

CodeUnet	
Input size	$T \times 3$
Embedding size	512
Transformer encoder layers	6
Transformer decoder layers every block	2
Temporal downsample size	4
Others	
Batch size per GPU	4
Optimizer	AdamW
Learning rate	3e-4
δ	0.01
λ	1.0

Table 2: Hyperparameters of Pose-VQVAE.

C. Training and Inference Algorithm

The whole training and inference algorithm is shown in Algorithm 1 and 2.

Algorithm 1 Training of the PoseVQ-Diffusion, given gloss sequence y, initial network parameters θ, loss weight λ and δ, learning rate η.

repeat
  $(y, s) \leftarrow$ sample training gloss-pose pair
  $x_0 \leftarrow$ Pose-VQVAE-Encoder(s)
  $c, \mathcal{L}_{\text{len}} \leftarrow$ TREnc.(y)
  $t \leftarrow$ Uniform$(\{1, ..., T\})$
  $x_t \leftarrow$ sample from $q(x_t|x_0)$ ▷ Eq.(3) and Eq.(5)
  $\mathcal{L} = \mathcal{L}_{ddm} + \mathcal{L}_{\text{len}}$ ▷ Eq.(10) and Eq. (13)
  $\theta \leftarrow \theta - \eta\nabla_\theta\mathcal{L}$
until converged

Algorithm 2 Inference of the PoseVQ-Diffusion, given gloss sequence y and its length M.

$c, \{L_1, .., L_M\} \leftarrow$ TREnc.(y)
init x_t with predicted length $\sum_{i=1}^{M} L_i$
$t \leftarrow T$
while do$t > 0$ do
  $x_t \leftarrow$ sample from $p_\theta(x_{t-1}|x_t, c)$ ▷ Eq.(8)
  $t \leftarrow t - 1$
end while PoseVQ-VAE-Decoder(x_t)

D. Results

In this section, we provide more visualization results. In Fig. 3, we show predicted sign pose sequences that are sampled every 2 frames for a total of 32 frames. Moreover, we provides some videos in additional mp4 files and more examples are shown in https://slpdiffusier.github.io/g2p-ddm.

Sign Gloss: DEUTSCH HIER LAND SUEDOST HOCH DRUCK FREUNDLICH

GT

Pose-VQVAE

PoseVQ-Diffusion

PTR

Sign Gloss: DIENSTAG HAUPTSAECHLICH SONNE ABER WOLKE AUCH HABEN2 WECHSELHAFT MOEGLICH REGEN ODER GEWITTER

GT

Pose-VQVAE

PoseVQ-Diffusion

PTR

Sign Gloss: DAZU LUFT ENORM WARM FEUCHT KOENNEN GEWITTER

GT

Pose-VQVAE

PoseVQ-Diffusion

PTR