Supplementary Materials for "Vector Quantized Diffusion Model with CodeUnet for Text-to-Sign Pose Sequences Generation"

Anonymous Author(s)
Affiliation
Address
email

1 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.

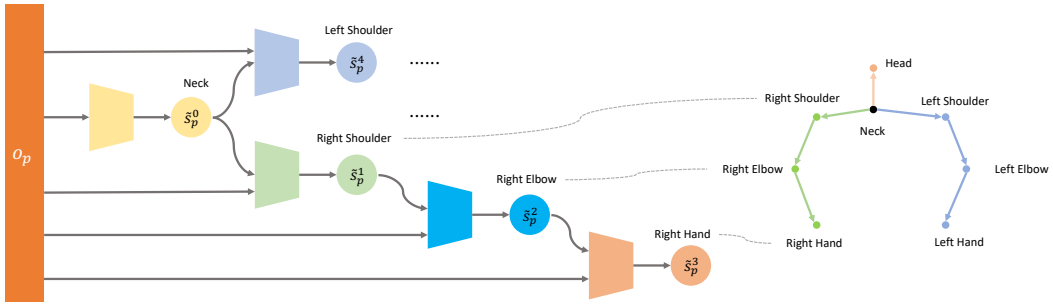

Figure 1: SPL for pose joints. Given the pose feature o_p, joint prediction $\tilde{s}_p^{(k)}$ are made hierarchically by following the spatial chain defined by the underlying skeleton.

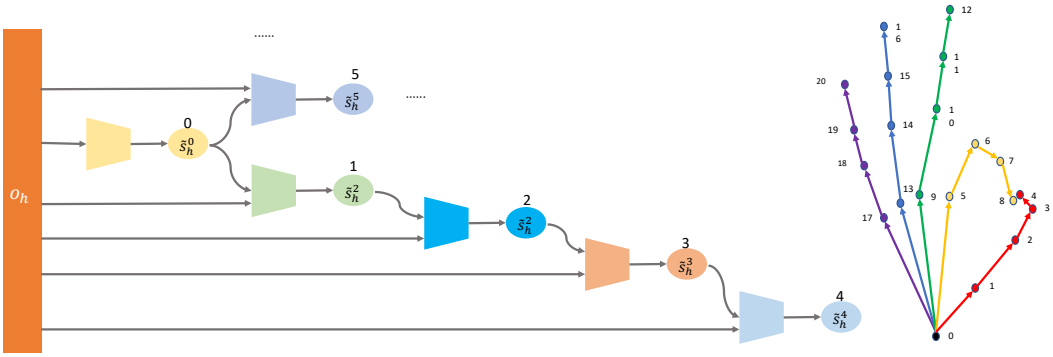

Figure 2: SPL for hand joints. Given the left hand feature or right hand feature o_h, joint prediction $\tilde{s}_h^{(k)}$ are made hierarchically by following the spatial chain defined by the underlying skeleton.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

2 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.

2.1 Pose-VQVAE

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	2048
Others	
Batch size per GPU	6
Learning rate	3e-4

Table 1: Hyperparameters of Pose-VQVAE.

2.2 PoseVQ-Diffusion

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
Learning rate	3e-4
δ	0.01
λ	1.0

Table 2: Hyperparameters of Pose-VQVAE.

3 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.

Figure 3: examples of predicted sign pose sequences compared with our reconstruction model and previous G2P model. For readability, we sampled every 2 frames for a total of 32 frames.