G2P-DDM: Generating Sign Pose Sequence from Gloss Sequence with Discrete Diffusion Model

Pan Xie[1], Qipeng Zhang[1], Peng Taiying[1], Hao Tang[2*], Yao Du[1], Zexian Li[1]

[1]Beihang University
[2]Carnegie Mellon University
{panxie, zhangqipeng, taiyi, duyaoo}@buaa.edu.cn, bjdxtanghao@gmail.com, lizexian0427@gmail.com

Abstract

The Sign Language Production (SLP) project aims to automatically translate spoken languages into sign sequences. Our approach focuses on the transformation of sign gloss sequences into their corresponding sign pose sequences (G2P). In this paper, we present a novel solution for this task by converting the continuous pose space generation problem into a discrete sequence generation problem. We introduce the Pose-VQVAE framework, which combines Variational Autoencoders (VAEs) with vector quantization to produce a discrete latent representation for continuous pose sequences. Additionally, we propose the G2P-DDM model, a discrete denoising diffusion architecture for length-varied discrete sequence data, to model the latent prior. To further enhance the quality of pose sequence generation in the discrete space, we present the CodeUnet model to leverage spatial-temporal information. Lastly, we develop a heuristic sequential clustering method to predict variable lengths of pose sequences for corresponding gloss sequences. Our results show that our model outperforms state-of-the-art G2P models on the public SLP evaluation benchmark. For more generated results, please visit our project page: https://slpdiffusier.github.io/g2p-ddm.

Introduction

Sign Language Production (SLP) is a crucial task for the Deaf community, involving the provision of continuous sign videos for spoken language sentences. Due to the distinct linguistic systems between sign languages and spoken languages (Pfau, Salzmann, and Steinbach 2018), sign languages have different sign orders, making direct alignment mapping between them challenging. To address this issue, prior works first translate spoken languages into glosses[1], followed by generating sign pose sequences based on the gloss sequences (G2P)(Saunders, Bowden, and Camgöz 2020; Saunders, Camgöz, and Bowden 2020). Finally, the generated sign pose sequence can optionally be used to produce a photo-realistic sign video(Saunders, Camgoz, and Bowden 2020). As such, G2P is the crucial procedure of this task, and this paper focuses on it.

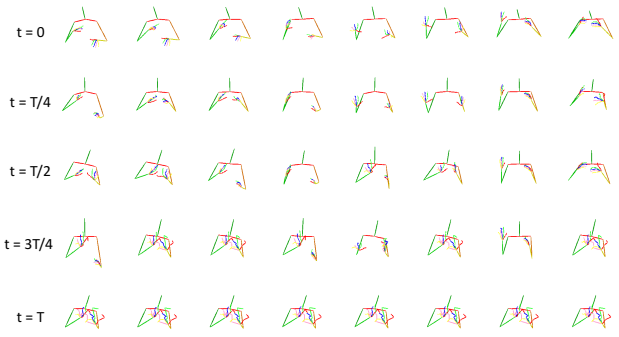

Figure 1: The forward diffusion process applied to a pose sequence. The first line (t=0) represents the original pose sequence. From top to bottom (t from 0 to T), the level of noise increases gradually.

Existing G2P methods can be broadly categorized as autoregressive (Saunders, Bowden, and Camgöz 2020; Saunders, Camgöz, and Bowden 2020) or non-autoregressive (Huang et al. 2021), depending on their decoding strategies. Autoregressive models generate the next pose frame based on previous frames, utilizing the teacher forcing strategy (Williams and Zipser 1989). However, during inference, recurrent decoding can lead to error propagation over time due to exposure bias (Schmidt 2019). To overcome this bottleneck, non-autoregressive methods have been proposed to enable the decoder to generate all target predictions simultaneously (Gu et al. 2018; Ghazvininejad et al. 2019). Huang *et al.* (Huang et al. 2021) introduced a non-autoregressive G2P model that generates sign pose sequences in a one-shot decoding scheme, using an External Aligner (EA) for sequence alignment learning.

Inspired by the remarkable results achieved by the recently developed Discrete Denoising Diffusion Probabilistic Models (D3PMs) (Hoogeboom et al. 2021; Austin et al. 2021; Gu et al. 2021) for language and vector quantized image generation, we propose a two-stage approach in this paper. Our method involves transforming the continuous pose sequence into discrete tokens and modeling the discrete prior space using the denoising diffusion architecture. The proposed method is an iterative non-autoregressive approach that performs parallel refinement on the generated results,

*Corresponding Author

[1]Sign glosses are minimal lexical items that match the meaning of signs and correspond to spoken language words.

demonstrating expressive generative capacity.

We elaborate our approach in three steps. Firstly, we represent the pose sequence as sequential latent codes using a vector quantized variational autoencoder (VQ-VAE). Unlike image VQ-VAE (Esser, Rombach, and Ommer 2021; van den Oord, Vinyals, and Kavukcuoglu 2017), we propose a specific architecture, Pose-VQVAE, that divides the sign skeleton into three local point patches representing pose, right hand, and left hand separately. Additionally, we use a multi-codebook to maintain separated latent embedding space for each local patch, resulting in stronger feature semantics. This approach eases the difficulty in constructing mappings between the sign pose feature and the codebook feature, thus improving reconstruction quality.

Next, we present G2P-DDM, which extends the standard discrete diffusion models (Austin et al. 2021; Gu et al. 2021) to model the sequential alignments between sign glosses and quantized codes of pose sequences. This approach employs a discrete diffusion model that samples the data distribution by reversing a forward diffusion process that gradually corrupts the input via a fixed Markov chain. The corruption process, depicted in Figure 1, achieved by adding noise data (*e.g.*, [MASK] token), draws our attention to the mask-based generative model, Mask-Predict (Ghazvininejad et al. 2019), which has been shown to be a variant of the diffusion model (Austin et al. 2021). We explore two variants of the diffusion model for variable-length sequence generation. To better leverage the spatial and temporal information of the quantized pose sequences, we introduce a new architecture, CodeUnet, which is a "fully transformer network" designed for discrete tokens. Through iterative refinements and improved spatial-temporal modeling, our model achieves a higher quality of conditional pose sequence generation.

Finally, we address the challenging task of length prediction in non-autoregressive G2P models, as the corresponding lengths of different sign glosses are variable. To tackle this issue, we propose a novel clustering method for this specific sequential data that local adjacent frames should belong to a cluster. Taking advantage of the meaningful learned codes in the first stage, we apply the k-nearest-neighbor-based density peaks clustering algorithm (Du, Ding, and Jia 2016; Zeng et al. 2022) to locate peaks with higher local density. We then design a heuristic algorithm to find the boundary between two peaks based on their semantic distance with the two peak codes. Finally, we leverage the length of each gloss as additional supervised information to predict the length of the gloss sequence during inference.

Our proposed model demonstrates significant improvement in the generation quality on the challenging RWTH-PHOENIX-WEATHER-2014T (Camgöz et al. 2018) dataset. The evaluation of conditional sequential generation is performed using a back-translated model. Extensive experiments show that our model increases the WER score from 82.01%(Huang et al. 2021) to 77.26% for the generated pose sequence to gloss sequence, and the BLEU score from 6.66(Huang et al. 2021) to 7.50 for the generated pose sequence to spoken language.

Related Works

Sign Language Production. Most sign language works focus on sign language recognition (SLR) and translation (SLT) (Camgöz et al. 2018, 2020; Camgöz et al. 2020; Zhou et al. 2022; Xie, Zhao, and Hu 2021; Hu et al. 2021), aiming to translate the video-based sign language into text-based sequences. And few attempts have been made for the more challenging task of sign language production (SLP) (Stoll et al. 2018; Xiao, Qin, and Yin 2020). Stoll *et al.* proposed the first deep SLP model, which adopts the three-step pipeline. In the core process for G2P, they learn the mapping between the sign glosses and the skeleton poses via a look-up table. After that, B. Saunders *et al.* (Saunders, Camgöz, and Bowden 2020) proposed the progressive transformer to learn the mapping with an encoder-decoder architecture and generate the sign pose in an autoregressive manner in the inference. Further, B. Saunders *et al.* (Saunders, Camgoz, and Bowden 2020) proposed a Mixture Density Network (MDN) to generate the pose sequences condition on the sign glosses and utilize a GAN-based method (Chan et al. 2019) to produce the photo-realistic sign language video. B. Saunders *et al.* (Saunders, Camgöz, and Bowden 2021) translated the spoken language to sign language representation with an autoregressive transformer network and used the gloss information to provide additional supervision. Then they proposed a Mixture of Motion Primitives(MoMP) architecture to combine distinct motion primitives to produce a continuous sign language sequence. B. Saunders *et al.* (Saunders, Camgoz, and Bowden 2022) propose a novel Frame Selection Network (FS-NET) to improve the temporal alignment of interpolated dictionary signs and SIGNGAN, a pose-conditioned human synthesis model that produces photo-realistic sign language videos direct from skeleton pose. Although they achieved state-of-the-art results, they used an additional sign language dictionary (Hanke et al. 2010), meaning that each sign vocabulary has a corresponding pose sequence. Therefore, this paper did not compare their results.

Different from these methods, Huang *et al.* (Huang et al. 2021) proposed a non-autoregressive model to parallelly generate the sign pose sequence avoiding the error accumulation problem. They applied the monotonic alignment search (Kim et al. 2020) to generate the alignment lengths of each gloss. Our model also explores a non-autoregressive method with a diffusion strategy, and the adopted diffusion model architecture allows us to refine the results with multiple iterations.

Discrete Diffusion Models. Most previous works focus on Gaussian diffusion processes that operate in continuous state spaces (Dhariwal and Nichol 2021; Ho, Jain, and Abbeel 2020; Ho et al. 2022; Nichol and Dhariwal 2021; Rombach et al. 2021). The discrete diffusion model is first introduced in (Sohl-Dickstein et al. 2015), and it is applied to text generation in Argmax Flow (Hoogeboom et al. 2021). To improve and extend the discrete diffusion model, D3PM (Austin et al. 2021) used a structured categorical corruption process to shape data generation and embed structure in the forward process. VQ-Diffusion (Gu et al. 2021) applied the discrete diffusion model to conditional vector quantized image synthesis with a mask-and-replace diffusion strategy. Upon this work, we extend this diffusion strategy with more special

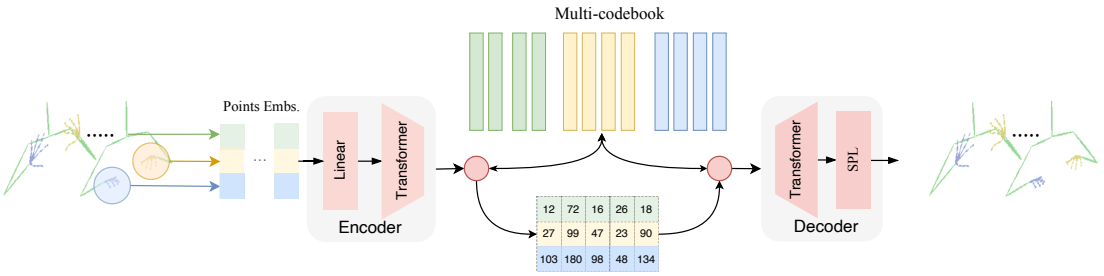

Figure 2: The architecture of the first stage model Pose-VQVAE for learning the discrete latent codes.

states to length-varied discrete sequence data and introduce an Unet-like "fully transformer" network to model spatial-temporal space.

The Proposed Method

Our paper aims to improve the generation of conditional sign pose sequences through an enhanced discrete diffusion model. Our approach consists of three key components: the Pose-VQVAE for latent code learning, the G2P-DDM with CodeUnet for prior learning to generate discrete codes, and a sequential-KNN algorithm for length prediction in a non-autoregressive approach.

Pose VQ-VAE

In this section, we introduce how to tokenize the points of a sign pose skeleton into a set of discrete tokens. A naive approach is to treat per point as one token. However, such a points-wise reconstruction model tends to have tremendous computational cost due to the quadratic complexity of self-attention in Transformers. On the other hand, since the details of hand points are essential for sign pose understanding, treating all the points into one token leads to remarkably inferior reconstruction performance. To achieve a better trade-off between quality and speed, we propose a simple yet efficient implementation that groups the points of a sign skeleton into three local patches, representing pose, right hand, and left hand separately. Figure 2 illustrates the framework of our proposed Pose-VQVAE model with the following submodules.

Encoder. Given a sign pose sequence of N frames $\mathbf{s} = (s_1, s_2, ..., s_n, ..., s_N) \in \mathbb{R}^{N\times J\times K}$, where $\{x_n^j\}_{j=1}^{J}$ presents a single sign skeleton containing J joints and K denotes the feature dimension for human joint data. We separate these points into three local paths, $\mathbf{s}_p \in \mathbb{R}^{N\times(J_p\times K)}$, $\mathbf{s}_r \in \mathbb{R}^{N\times(J_r\times K)}$, and $\mathbf{s}_l \in \mathbb{R}^{N\times(J_l\times K)}$ for the pose, right hand, and left hand, respectively, where $J = J_p + J_r + J_l$. In the encoder module $E(e|\mathbf{s})$, we first transform these three point sequences into feature sequences by simple three linear layers and concatenate them together. Then we apply a spatial-temporal Transformer network to learn the long-range interactions within the sequential point features. Finally, we arrive at the encoded features $\{e_n \in \mathbb{R}^{3\times h}\}_{n=1}^{N}$.

Multi-Codebook. Similar to image VQ-VAE (van den Oord, Vinyals, and Kavukcuoglu 2017), we take the encoded features as inputs and convert them into discrete tokens. Specifically, we perform the nearest neighbors method $\mathcal{Q}(z|e)$ to quantize the point feature to the quantized features $\{z_n \in \mathbb{R}^{3\times h}\}_{n=1}^{N}$. The quantized features are maintained by three separate codebooks, where each codebook is of size V.

Decoder. The decoder $D(\tilde{\mathbf{s}}|z)$ receives the quantized features as inputs and also applies spatial-temporal Transformer to get the output features $\{o_n \in \mathbb{R}^{3\times h}\}_{n=1}^{N}$. Finally, we separate the output feature for three sub-skeleton and utilize a structured prediction layer (SPL) (Aksan, Kaufmann, and Hilliges 2019) $\mathcal{P}(\tilde{s}|o)$ to reconstruct the corresponding sub-skeleton $\tilde{\mathbf{s}}_p \in \mathbb{R}^{N\times(J_r\times K)}$, $\tilde{\mathbf{s}}_l \in \mathbb{R}^{N\times(J_r\times K)}$, and $\tilde{\mathbf{s}}_r \in \mathbb{R}^{N\times(J_r\times K)}$. We adopt the SPL to rebuild the skeleton from feature because it explicitly models the spatial structure of the human skeleton and the spatial dependencies between joints. The hierarchy chains of the pose, right hand, and left hand skeleton are given in the Appendix.

Training. The encoder $E(e|\mathbf{s})$, tokenizer $\mathcal{Q}(z|e)$, and decoder $D(\tilde{\mathbf{s}}|z)$ can be trained end-to-end via the following loss function:

$$\begin{aligned}\mathcal{L}_{\text{Pose-VQVAE}} =& ||\mathbf{s}_p - \tilde{\mathbf{s}}_p|| + ||\mathbf{s}_r - \tilde{\mathbf{s}}_r|| + ||\mathbf{s}_l - \tilde{\mathbf{s}}_l|| + \\ & ||sg[e] - z|| + \beta||sg[z] - e||,\end{aligned} \tag{1}$$

where $sg[\cdot]$ stands for stop-gradient operation.

G2P-DDM with CodeUnet

To allow conditional sampling, a discrete diffusion model is trained on the latent codes obtained from the Pose-VQVAE model. Figure 3 shows the architecture of our proposed G2P-DDM, which aims to model the latent space in an iterative non-autoregressive manner.

Given a sequence of latent codes $x_0 \in \mathbb{R}^{N\times 3}$ obtained from the vector quantized model, where $x_0^{(i,j)} \in \{1, 2, ..., V\}$ at location (i, j) represents the index within the codebook. The diffusion process aims to corrupt the original data x_0 via a fixed Markov chain $p(x_t|x_{t-1})$ by adding a small amount of noise continuously. After a fixed T timesteps, it produces a sequence of increasingly noisy data $x_1, .., x_T$ with the same dimensions as x_0, and x_T becomes a pure noise sample.

For the scalar discrete variables with V categories $x_t^{(i,j)} \in [1, V]$, the forward transition probabilities from x_{t-1} to x_t can be represented by matrices $[Q_t]_{mn} = q(x_t = m|x_{t-1} = n) \in \mathbb{R}^{V\times V}$. Note that we omit the superscripts (i, j) to avoid confusion. Then the forward diffusion process can be written as:

$$q(x_t|x_{t-1}) = \mathbf{x}_t^T Q_t \mathbf{x}_{t-1}, \tag{2}$$

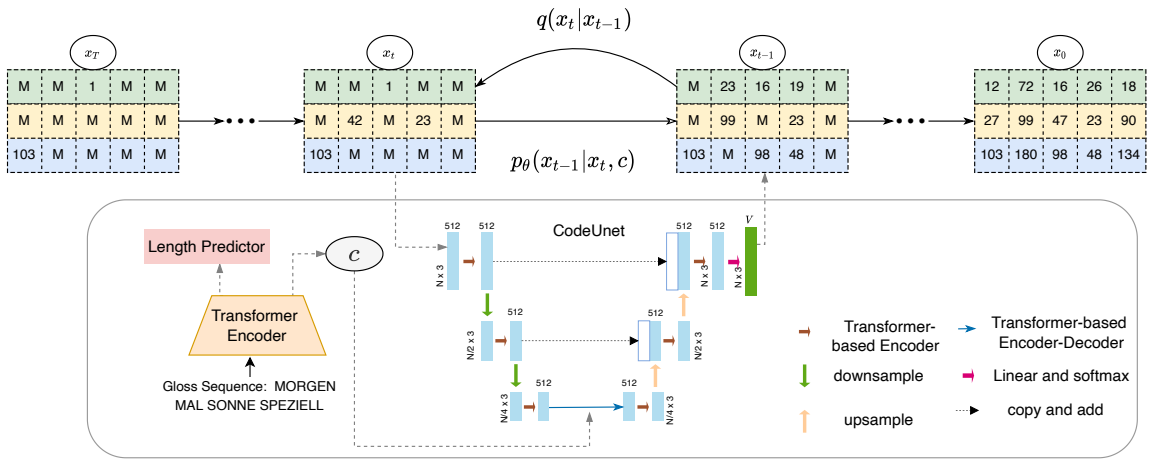

Figure 3: Our approach uses a discrete diffusion model to represent the conditional sign pose sequence generation. Specifically, each quantized code is randomly masked or replaced, and a CodeUnet model is trained to restore the original data.

where $\mathbf{x}_t \in \mathbb{R}^{V\times 1}$ is the one-hot version of x_t and $Q_t\mathbf{x}_{t-1}$ is the categorical distribution for x_t. A nice property of the above Markov diffusion process is that we can sample x_t as any timestep directly from x_0 as:

$$q(x_t|x_0) = \mathbf{x}_t^T \bar{Q}_t \mathbf{x}_0, \text{with } \bar{Q}_t = Q_t \dots Q_1. \tag{3}$$

D3PM (Austin et al. 2021) formulates the transition matrix $Q_t \in \mathbb{R}^{V\times V}$ by introducing a small number of uniform noises to the categorical distribution. Based on D3PM, VQ-Diffusion (Gu et al. 2021) proposes a mask-and-replace diffusion strategy that not only replaces the previous value but also inserts [MASK] token to explicitly figure out the tokens that have been replaced. We extend this mask-and-replace strategy to our variable-length sequence modeling. Since the length of pose sequences may be different in a minibatch, we have to add two special tokens, [MASK] and [PAD] tokens, so each token has $V+2$ states. The mask-and-replace diffusion process can be defined as follows: each token has a probability of α_t to be unchanged, $V\beta_t$ to be uniformly resampled, and $\gamma_t = 1-\alpha_t - V\beta_t$ to be replaced with [MASK] token. Note that [MASK] and [PAD] tokens always keep their own state. The transition matrix $Q_t \in \mathbb{R}^{(V+2)\times(V+2)}$ is formulated as the second matrix of the following:

$$Q_t = \begin{bmatrix} \alpha_t+\beta_t & \beta_t & \cdots & \beta_t & 0 & 0 \\ \beta_t & \alpha_t+\beta_t & \cdots & \beta_t & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \beta_t & \beta_t & \cdots & \alpha_t+\beta_t & 0 & 0 \\ \gamma_t & \gamma_t & \cdots & \gamma_t & 1 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 \end{bmatrix}. \tag{4}$$

Finally, the categorical distribution of $\mathbf{x}_t$ can be derived as following using reparameterization trick:

$$\text{when } x_0 \neq V+2, \quad \bar{Q}_t\mathbf{x}_0 = \begin{cases} \bar{\alpha}_t + \bar{\beta}_t, & x_t = x_0 \\ \bar{\beta}_t, & x_t \neq x_0 \text{ and } x_t \leq V \\ \bar{\gamma}_t, & x_t = V+1 \\ 0, & x_t = V+2 \end{cases}$$
$$\text{when } x_0 = V+2, \quad \bar{Q}_t\mathbf{x}_0 = \begin{cases} 0, & x_t \neq V+2 \\ 1, & x_t = V+2 \end{cases} \tag{5}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$, $\bar{\gamma}_t = 1 - \prod_{i=1}^{t}(1-\gamma_i)$, and $\bar{\beta}_t = (1-\bar{\alpha}_t - \bar{\gamma}_t)/V$. Therefore, we can directly sample x_t within the computation cost $O(V)$. A visualized example of the diffusion process is shown in Figure 1, we first get the noised latent codes by $q(x_t|x_t)$ and decode them to sign skeleton with Pose-VQVAE decoder module.

The reverse denoising process is similar to D3PM (Austin et al. 2021) and VQ-Diffusion (Gu et al. 2021). The relevant derivation process is given in the appendix.

CodeUnet for Model Learning. Most image diffusion models (Dhariwal and Nichol 2021; Ho, Jain, and Abbeel 2020; Song et al. 2021) adopt the Unet (Ronneberger, Fischer, and Brox 2015) as their architectures since it is effective for data with spatial structure. However, directly applying the Unet in discrete sequence generation, *e.g.*, text generation (Austin et al. 2021) and quantized image synthesis (Gu et al. 2021), will bring information leakage problem since the convolution layer over adjacent tokens may provide shortcuts for the mask-based prediction (Nawrot et al. 2021). Therefore, Austin *et al.* (Austin et al. 2021) and Gu *et al* (Gu et al. 2021) used the token-wise Transformer framework to learn the distribution $p_\theta(\tilde{x}_0|x_t, c)$. In this work, to incorporate the advantages of Unet and Transformer networks, we propose a novel architecture, CodeUnet, to learn the spatial-temporal interaction for our quantized pose sequence generation.

As shown in Figure 3, the CodeUnet consists of a contracting path (left side), an expansive path (right side), and a middle module. The middle module is an encoder-decoder Transformer framework. The encoder consists of 6 Trans-

former blocks. It takes the gloss sentence as input and obtains a conditional feature sequence. The decoder has two blocks. Each block has a self-attention, a cross-attention, a feed-forward network, and an Adaptive Layer Normalization (AdaLN) (Ba, Kiros, and Hinton 2016; Gu et al. 2021). The AdaLN operator is devised to incorporate timestep t information as $\text{AdaLN}(h, t) = \alpha_t \text{LayerNorm}(h) + \beta_t$, where h is the intermediate activations, α_t and β_t are obtained from a linear projection of the timestep embedding.

The contracting and expansive paths are hierarchical structures, and each level has two Transformer encoder blocks. For downsampling in contracting path, given the feature of quantized pose sequence, *e.g.*, $h \in \mathbb{R}^{N\times 3\times d_{\text{model}}}$, where d_{model} is the feature dimension, we first sample uniformly with stride 2 in the temporal dimension and remain constant in the spatial dimension. Then we set the downsampled feature as query $Q \in \mathbb{R}^{N/2\times 3\times d_{\text{model}}}$, and keep key K and value V unchanged for the following attention network. In the upsampling of the expansive path, we directly repeat the feature 2 times as a query, but the key and value remain for the following attention network:

$$\forall n = 1, ..., N, Q_n^{\text{up}} = h_{n//2}, K^{\text{up}} = V^{\text{up}} = h, \quad (6)$$

where $\cdot//\cdot$ denotes floor division. Finally, a linear layer and a softmax layer are applied to make the prediction.

Length Prediction with Sequential-KNN

In this section, inspired by (Zeng et al. 2022), which merges tokens with similar semantic meanings from different locations, we propose a novel clustering algorithm to get the lengths for corresponding glosses. Specifically, given a token sequence that is obtained from the Pose-VQVAE model, we compute the local density ρ of each token according to its k-nearest-neighbors:

$$\rho_i = exp(-\frac{1}{k}\sum_{z_j \in \text{KNN}(z_i)} \|z_i - z_j\|_2^2), \text{ where } |i-j| <= l \quad (7)$$

where i, j is the position in the sequence, and l is a predefined hyperparameter indicating that we only consider the local region since the adjacent tokens are more likely to belong to a gloss. z_i and z_j are the latent feature for i^{th} and j^{th} tokens. $\text{KNN}(x_i)$ represents the k-nearest neighbors for i^{th} token.

We assign $\{p_1, ..., p_M\}$ positions with a higher local density as the peaks, where M is the length of the gloss sequence. Then between two adjacent peaks, for example, p_1 and p_2, we sequentially iterate from p_1 to p_2 and find the first position that is farther from z_{p_1} and closer to z_{p_2}, which is the boundary we determined. After finding these boundaries, we get the lengths of the contiguous pose sequence for its corresponding glosses. As shown in Figure 3, we define the obtained lengths as $\{L_1, .., L_M\}$, and the Transformer encoder for gloss sequence is trained under the supervised information of lengths. For each gloss word, we predict a number from $[1, P]$, where P is the maximum length of the target pose sequence. Mathematically, we formulate the classification loss of length prediction as:

$$\mathcal{L}_{\text{len}} = \frac{\delta}{M}\sum_{i}^{M}\sum_{j}^{P}(-L_i = j)\log p(L_i|c). \quad (8)$$

In the training of the discrete diffusion mode, $\mathcal{L}_{\text{len}}$ is trained together with a coefficient δ. In the inference, we predict the length of glosses, and their summation is the length of the target pose sequence.

In summary, we arrive at our proposed two-stage approach, G2P-DDM, with the first-stage Pose-VQVAE model and the second-stage discrete diffusion model with a length predictor.

Experiments

Datasets. We evaluate our G2P model on RWTH-PHOENIX-WEATHER-2014T dataset (Camgöz et al. 2018). It is the *only* publicly available SLP dataset with parallel sign language videos, gloss annotations, and spoken language translations. This corpus contains 7,096 training samples (with 1,066 different sign glosses in gloss annotations and 2,887 words in German spoken language translations), 519 validation samples, and 642 test samples.

Evaluation Metrics. Following the widely-used setting in SLP (Saunders, Camgöz, and Bowden 2020), we adopt the back-translation method for evaluation. Specifically, we utilize the state-of-the-art SLT (Camgöz et al. 2020) model to translate the generated sign pose sequence back to gloss sequence and spoken language, where its input is modified as pose sequence. Specifically, we compute BLEU (Papineni et al. 2002) and Word Error Rate (WER) between the back-translated spoken language translations and gloss recognition results with ground truth spoken language and gloss sequence. Although this evaluation method may introduce noise, it is currently the prevailing approach in SLP models, and we adopt it to ensure a fair comparison with existing methods.

Data Processing. Since the RWTH-PHOENIX-WEATHER-2014T dataset does not contain pose information, we generate the pose sequence as the ground truth. Following B. Saunders *et al.* (Saunders, Camgöz, and Bowden 2020), we extract 2D joint points from sign video using OpenPose (Cao et al. 2021) and lift the 2D joints to 3D with a skeletal model estimation improvement method (Zelinka and Kanis 2020). Finally, similar to (Stoll et al. 2018), we apply skeleton normalization to remove the skeleton size difference between different signers.

Model Settings. The Pose-VQVAE consists of an Encoder, a Tokenizer, and a Decoder. The Encoder contains a linear layer to transform pose points to a hidden feature with a dimension set as 256, a 3-layer Transformer module with divided space-time attention (Bertasius, Wang, and Torresani 2021). The Tokenizer maintains a codebook with a size set as 2,048. The Decoder contains the same 3-layer Transformer module as the Encoder and an SPL layer to predict the structural sign skeleton. For the discrete diffusion model, we set the timestep T as 100. All Transformer blocks of CodeUnet have d_{model}=512 and N_{depth}=2. The size of the local region l in Eq. (7), is set as 16, which is the average length of a gloss. And the number of nearest neighbors k is set as 16. We train the model on 8 NVIDIA Tesla V100 GPUs. We include all hyperparameters settings and the details of implementation in the Appendix.

Method	WER	BLEU-1	BLEU-2	BLEU-3	BLEU-4	DTW-MJE
PTR† (Saunders, Camgöz, and Bowden 2020)	94.65	11.45	7.08	5.08	4.04	0.191
NAT-AT (Huang et al. 2021)	88.15	14.26	9.93	7.11	5.53	0.177
NAT-EA (Huang et al. 2021)	82.01	15.12	10.45	7.99	6.66	0.146
G2P-AR (Ours)	85.27	14.26	10.02	7.57	5.94	0.172
G2P-MP (Ours)	79.38	15.43	10.69	8.26	6.98	0.146
G2P-DDM (Ours)	**77.26**	**16.11**	**11.37**	**9.22**	**7.50**	**0.116**
GT†	55.93	24.12	16.77	12.80	10.58	0.0

Table 1: Quantitative results for the G2P task on the RWTH-PHOENIX-WEATHER-2014T test dataset. † indicates the results is provided by Huang et al. (Huang et al. 2021). Note that, GT refers to the validation metrics obtained by using the original pose sequence extracted from the video and then applying a back-translation method.

Comparisons with State-of-the-Art Methods

Competing Methods. We compare our G2P-DDM with previous state-of-the-art G2P models. **Progressive Transformer** (PTR) (Saunders, Camgöz, and Bowden 2020) is the first SLP model to tackle the G2P problem in an autoregressive manner. Since they use the ground-truth first sign pose frame and timing information, their reported results are not comparable to ours. Thus we adopt the results reported by Huang *et al.* (Huang et al. 2021). **NAT-EA** (Huang et al. 2021) proposes a non-autoregressive method to directly predict the target pose sequence with the External Aligner (EA) to learn alignments between glosses and pose sequences. **NAT-AT** is the NAT model without EA that uses the decoder-to-encoder attention to learn the alignments.

Quantitative Comparison. The comparison between our G2P-DDM and the competing methods is shown in Tabel 1. Note that, the evaluation results of the GT† are lower than the reported results in the state-of-the-art SLT (Camgöz et al. 2020) model. This is because the evaluation results obtained using the pose sequence are inferior to those obtained using photo-realistic content (Saunders, Camgoz, and Bowden 2022).

The row of **G2P-AR** refers to the vector quantized model with an autoregressive decoder. The row of **G2P-MP** refs to the vector quantized model with the Mask-Predict (Ghazvininejad et al. 2019) strategy, which is also a variant of discrete diffusion model (Austin et al. 2021). **G2P-DDM** refs to the vector quantized model with mask-and-replace diffusion strategy. As indicated in Table 1, both diffusion-based models outperform the state-of-the-art G2P models with relative improvements on WER score by 5.7% (82.01 $\rightarrow$ 77.26) and on BLEU-4 by 12.6% (6.66 $\rightarrow$ 7.50). This shows the effectiveness of the iterative mask-based non-autoregressive method on the vector quantized pose sequence. In addition, the Mask-Predict strategy is a mask-only strategy similar to G2P-DDM with $\bar{\gamma}_T = 1$. Therefore, G2P-DDM achieves better performance than G2P-MP. This reflects that the mask-and-replace strategy is superior to the mask-only strategy.

Model Analysis and Discussions

We also investigate the effects of different components and design choices of our proposed model.

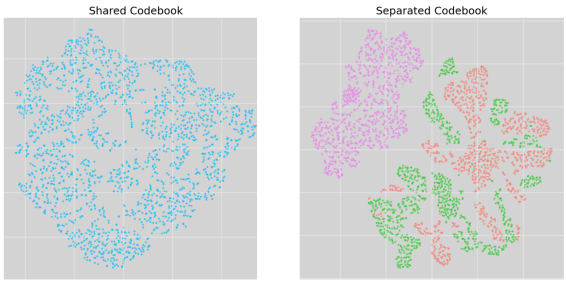

Figure 4: Visulization of latent vectors in the shared codebook and separated codebooks. In the separated codebook, the pink part is for the pose, and the green and orange parts represent the left and right hands, respectively.

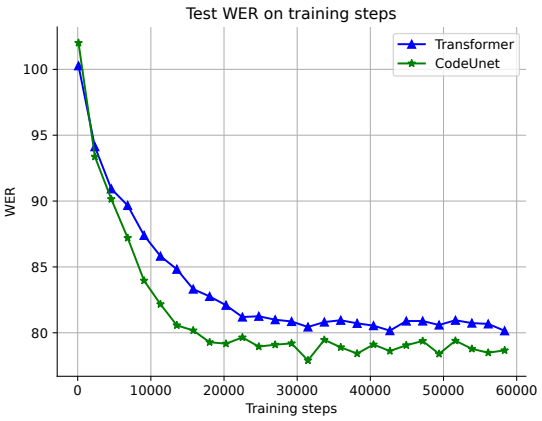

Figure 5: Ablation on the design of prediction model.

Analysis of The Design of Pose-VQVAE. As shown in Table 2, we study the design of our Pose-VQVAE model. Pose-VQVAE-joint-shared means we compress all points into one token with one shared codebook. Pose-VQVAE-separated-shared means the points are separated into three local patches according to the structure of a sign skeleton, and the latent embedding space is maintained with one shared codebook. Pose-VQVAE-separated-separated means the points are separated into three local patches, and the latent vectors are maintained with three codebooks separately.

Experimental results in Table 2 show that Pose-VQVAE-

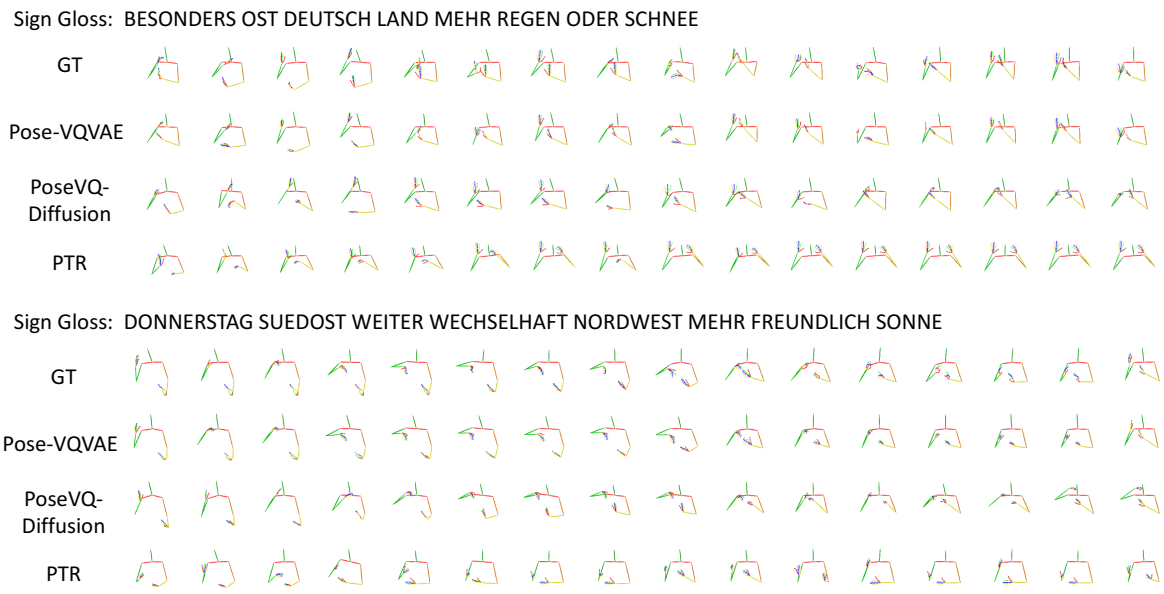

Figure 6: G2P qualitative results. We show several examples of generated sign pose sequences compared with Pose-VQVAE and previous G2P model (Saunders, Camgöz, and Bowden 2020). For readability, we sampled every 5 frames for a total of 16 frames. See the Appendix for more results.

local patches	codebook (size)	MSE (↓)	WER (↓)
joint	shared (2048)	0.0242	-
separated	shared (2048)	0.0139	78.21
separated	shared (3072)	0.0131	78.15
separated	separated (1024*3)	**0.0113**	**77.26**

Table 2: Ablation on the design of Pose-VQVAE reconstruction model.

		Training Steps			
		20	50	100	200
Infer. Steps	20	79.53	79.40	78.25	78.62
	50	-	79.31	77.69	78.23
	100	-	-	**77.26**	78.18
	200	-	-	-	78.15

Table 3: Ablation on training steps and inference steps.

separated-separated achieves much better reconstruction (MSE) performance. This indicates that compressing all skeleton points into one token embedding is not advisable, leading to information loss. Using separated latent feature spaces for different local regions, that is, three codebooks can achieve better reconstruction quality and generation performance. To further explain this phenomenon, we visualize the latent space vectors of shared codebooks and separated codebooks with T-SNE (Van der Maaten and Hinton 2008). As shown in Figure 4, the latent space vectors corresponding to the left-hand and right-hand local regions are easily confused because of their close distances. Therefore, separated codebooks can reduce the difficulty in constructing mappings between the sign pose feature and the codebook feature, thus learning better latent space and reconstruction quality. The second row of Figure 6 shows the sample of sign pose sequences reconstructed by Pose-VQVAE-separated-separated.

CodeUnet vs. Transformer. For a fair comparison, we replace our CodeUnet with a Transformer network, keeping other settings the same. As shown in Figure 5, the diffusion-based model with our CodeUnet achieves better performance on the back-translate evaluation. This phenomenon suggests that the hierarchical structure of CodeUnet makes it particularly effective for data with spatial structure. Moreover, the curve in the figure shows that CodeUnet coverages faster than Transformer. Having said that, due to sign pose sequences being temporally redundant, the compression of CodeUnet in the time dimension makes it more efficient in training.

Number of Timesteps. We compare the performance of the model with different numbers of training steps. As shown in the left two columns of Table 3, we find that the results get better when the training step size is increased from 20 to 100. As it increased further, the results seemed to saturate. Therefore, we set the training step to 100 to trade off performance and speed. Besides, within the same training steps, increasing the number of inference steps yields better results.

Deaf User Evaluation In our final user evaluation, we provided 50 pose sequences generated by our proposed method and a baseline method (Saunders, Camgöz, and Bowden 2020), and asked 7 participants to compare which one was closer to the ground truth pose sequence. The results showed that 319/350 preferred our method, while only 31/350 chose the baseline method. This clearly demonstrates the superiority of our proposed approach.

Conclusion

We present a novel paradigm for text-based sign pose sequence generation. Specifically, we first devise a specific architecture Pose-VQVAE with a multi-codebook to learn

semantic discrete codes by reconstruction. Then we extend the discrete diffusion method with special states to model the alignments between sign glosses and length-varied quantized code sequences. Further, a "fully transformer" network CodeUnet is proposed to model the spatial-temporal information in discrete space. Finally, we propose a sequential-KNN algorithm to learn the length of corresponding glosses and then predict the length as a classification task. Our extensive experiments show that our proposed G2P-DDM framework outperforms previous state-of-the-art methods.

References

Aksan, E.; Kaufmann, M.; and Hilliges, O. 2019. Structured Prediction Helps 3D Human Motion Modelling. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 7143–7152. IEEE.

Austin, J.; Johnson, D. D.; Ho, J.; Tarlow, D.; and van den Berg, R. 2021. Structured Denoising Diffusion Models in Discrete State-Spaces. In *NeurIPS*.

Ba, J.; Kiros, J. R.; and Hinton, G. E. 2016. Layer Normalization. *ArXiv preprint*, abs/1607.06450.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is Space-Time Attention All You Need for Video Understanding? In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 813–824. PMLR.

Camgöz, N. C.; Hadfield, S.; Koller, O.; Ney, H.; and Bowden, R. 2018. Neural Sign Language Translation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 7784–7793. IEEE Computer Society.

Camgöz, N. C.; Koller, O.; Hadfield, S.; and Bowden, R. 2020. Multi-channel Transformers for Multi-articulatory Sign Language Translation. *ArXiv preprint*, abs/2009.00299.

Camgöz, N. C.; Koller, O.; Hadfield, S.; and Bowden, R. 2020. Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 10020–10030. IEEE.

Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2021. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43: 172–186.

Chan, C.; Ginosar, S.; Zhou, T.; and Efros, A. A. 2019. Everybody Dance Now. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 5932–5941. IEEE.

Dhariwal, P.; and Nichol, A. 2021. Diffusion Models Beat GANs on Image Synthesis. In *NeurIPS*.

Du, M.; Ding, S.; and Jia, H. 2016. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowl. Based Syst.*, 99: 135–145.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming Transformers for High-Resolution Image Synthesis. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12868–12878.

Ghazvininejad, M.; Levy, O.; Liu, Y.; and Zettlemoyer, L. 2019. Mask-Predict: Parallel Decoding of Conditional Masked Language Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6112–6121. Hong Kong, China: Association for Computational Linguistics.

Gu, J.; Bradbury, J.; Xiong, C.; Li, V. O. K.; and Socher, R. 2018. Non-Autoregressive Neural Machine Translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Yuan, L.; and Guo, B. 2021. Vector Quantized Diffusion Model for Text-to-Image Synthesis. *ArXiv preprint*, abs/2111.14822.

Hanke, T.; König, L.; Wagner, S.; and Matthes, S. 2010. DGS Corpus & Dicta-Sign: The Hamburg Studio Setup. In *sign-lang@ LREC 2010*, 106–109. European Language Resources Association (ELRA).

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ho, J.; Saharia, C.; Chan, W.; Fleet, D.; Norouzi, M.; and Salimans, T. 2022. Cascaded Diffusion Models for High Fidelity Image Generation. *J. Mach. Learn. Res.*, 23: 47:1–47:33.

Hoogeboom, E.; Nielsen, D.; Jaini, P.; Forr'e, P.; and Welling, M. 2021. Argmax Flows and Multinomial Diffusion: Towards Non-Autoregressive Language Models. *ArXiv preprint*, abs/2102.05379.

Hu, H.; Zhao, W.; gang Zhou, W.; Wang, Y.; and Li, H. 2021. SignBERT: Pre-Training of Hand-Model-Aware Representation for Sign Language Recognition. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 11067–11076.

Huang, W.; Pan, W.; Zhao, Z.; and Tian, Q. 2021. Towards Fast and High-Quality Sign Language Production. *Proceedings of the 29th ACM International Conference on Multimedia*.

Kim, J.; Kim, S.; Kong, J.; and Yoon, S. 2020. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Nawrot, P.; Tworkowski, S.; Tyrolski, M.; Kaiser, L.; Wu, Y.; Szegedy, C.; and Michalewski, H. 2021. Hierarchical Transformers Are More Efficient Language Models. *ArXiv preprint*, abs/2110.13711.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved Denoising Diffusion Probabilistic Models. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 8162–8171. PMLR.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Pfau, R.; Salzmann, M.; and Steinbach, M. 2018. The syntax of sign language agreement: Common ingredients, but unusual recipe. *Glossa: a journal of general linguistics*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *ArXiv preprint*, abs/2112.10752.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*.

Saunders, B.; Bowden, R.; and Camgöz, N. C. 2020. Adversarial Training for Multi-Channel Sign Language Production. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press.

Saunders, B.; Camgoz, N. C.; and Bowden, R. 2020. Everybody Sign Now: Translating Spoken Language to Photo Realistic Sign Language Video. *ArXiv preprint*, abs/2011.09846.

Saunders, B.; Camgöz, N. C.; and Bowden, R. 2020. Progressive Transformers for End-to-End Sign Language Production. *ArXiv preprint*, abs/2004.14874.

Saunders, B.; Camgöz, N. C.; and Bowden, R. 2021. Mixed SIGNals: Sign Language Production via a Mixture of Motion Primitives. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 1899–1909. IEEE.

Saunders, B.; Camgoz, N. C.; and Bowden, R. 2022. Signing at Scale: Learning to Co-Articulate Signs for Large-Scale Photo-Realistic Sign Language Production. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5141–5151.

Schmidt, F. 2019. Generalization in Generation: A closer look at Exposure Bias. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, 157–167. Hong Kong: Association for Computational Linguistics.

Sohl-Dickstein, J.; Weiss, E. A.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In Bach, F. R.; and Blei, D. M., eds., *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, 2256–2265. JMLR.org.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Stoll, S.; Camgöz, N. C.; Hadfield, S.; and Bowden, R. 2018. Sign Language Production using Neural Machine Translation and Generative Adversarial Networks. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, 304. BMVA Press.

van den Oord, A.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural Discrete Representation Learning. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 6306–6315.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Williams, R. J.; and Zipser, D. 1989. A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. *Neural Computation*, 1: 270–280.

Xiao, Q.; Qin, M.; and Yin, Y. 2020. Skeleton-based Chinese sign language recognition and generation for bidirectional communication between deaf and hearing people. *Neural networks : the official journal of the International Neural Network Society*, 125: 41–55.

Xie, P.; Zhao, M.; and Hu, X. 2021. PiSLTRc: Position-informed Sign Language Transformer with Content-aware Convolution. *ArXiv preprint*, abs/2107.12600.

Zelinka, J.; and Kanis, J. 2020. Neural Sign Language Synthesis: Words Are Our Glosses. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 3384–3392.

Zeng, W.; Jin, S.; Liu, W.; Qian, C.; Luo, P.; Wanli, O.; and Wang, X. 2022. Not All Tokens Are Equal: Human-centric Visual Analysis via Token Clustering Transformer. *ArXiv preprint*, abs/2204.08680.

Zhou, H.; gang Zhou, W.; Zhou, Y.; and Li, H. 2022. Spatial-Temporal Multi-Cue Network for Sign Language Recognition and Translation. *IEEE Transactions on Multimedia*, 24: 768–779.