2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

A Gloss-free Sign Language Production with Discrete Representation

Eui Jun Hwang, Huije Lee, and Jong C. Park[†]
Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea

***Abstract*— Gloss-free Sign Language Production (SLP) offers a direct translation of spoken language sentences into sign language, bypassing the need for gloss intermediaries. Previous autoregressive SLP methods have not fully achieved true autoregression, as they often depend on ground-truth data during inference. To fill this gap, we introduce Sign language Vector Quantization Network (SignVQNet), leveraging discrete spatio-temporal representations of sign poses. With such a discrete representation, our method incorporates beam search, a decoding strategy widely used in Natural Language Processing. Furthermore, we align the discrete representation with linguistic features from pre-trained language models such as BERT. Our results show the superior performance of our method over prior SLP methods in generating accurate and realistic sign pose sequences. Additionally, our analysis shows that the reliability of Back-Translation and Fréchet Gesture Distance as evaluation metrics, in contrast to DTW-MJE. The code and models are available at https://github.com/eddie-euijun-hwang/SignVQNet.**

I. INTRODUCTION

Gloss-free Sign Language Production (SLP) directly translates spoken language into sign poses, eliminating the need for gloss annotation. Glosses, while providing a direct mapping between spoken language and sign poses, require significant labor, time, and specialized knowledge of sign language. This high demand for the resources has been a driving factor in the growing interest and transition towards gloss-free methods [17], [18], [27], [31]. While the gloss-free methods typically exhibit lower performance than the gloss-based ones, they offer greater accessibility and efficiency.

In the domain of gloss-free SLP, two approaches prevail: retrieval and generative models. Retrieval models [6], [9], [26] fetch relevant samples from datasets based on textual prompts. Generative models [13], [24], [25], on the other hand, can generate entirely new signing sequences by leveraging patterns learned during training. This capability to produce diverse outputs makes generative models a compelling choice for SLP, which is the focus of our research. However, the generative models face a few challenges. The length disparity between sign pose sequences and their spoken equivalents often necessitates clustering sequences into gloss-level representations [18]. Moreover, the non-linear nature of sign language compared to the linear structure of spoken language adds complexity to this task.

Recent studies [12], [13] have pointed out the constraints of the model introduced earlier [24]. A significant concern is about its dependence on the initial ground-truth pose and timing for inference, pivotal for the model's autoregression. The continuous nature of the sign pose sequences, represented as keypoint data, complicates achieving true autoregression without auxiliary information during inference.

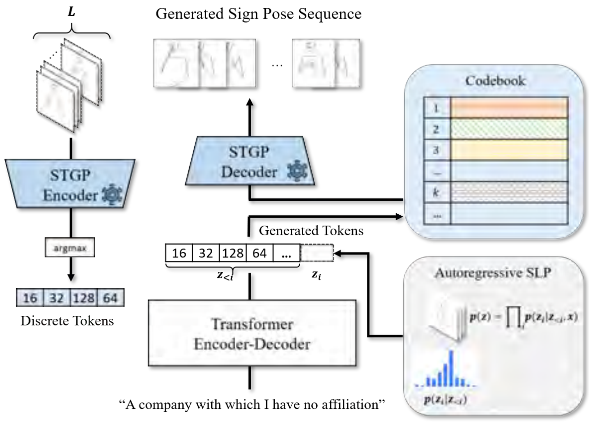


Fig. 1: An overview of SignVQNet, where both the pre-trained STGP encdoer and decoder remain frozen. The encoder converts the sign pose sequence into discrete tokens. These tokens are then generated from textual inputs by Transformer. The decoder transforms the generated tokens back into an actual sign pose sequence.

To address this, we introduce **Sign** language **V**ector **Q**uantization **Net**work (**SignVQNet**), as shown in Fig. 1. Leveraging vector quantization, our method converts the sign pose sequences into discrete tokens, enabling genuine autoregressive generation. This approach also supports beam search, commonly used for Natural Language Processing (NLP) tasks [7], [15], [16]. Additionally, we introduce latent-level alignment to directly associate linguistic features with sign pose features.

We compared the performance of SignVQNet against those of other existing SLP models using Back-Translation (BT) [24], DTW-MJE [12], and Fréchet Gesture Distance (FGD) [32]. Our experimental results showed that SignVQNet consistently outperformed the previous methods on two sign language datasets: RWTH-PHOENIX-WEATHER-2014T [3] and How2Sign [10]. In our additional experiments focusing on beam size adjustment of our method, we found that DTW-MJE suffers from inconsistencies, raising questions about its reliability as a suitable metric for SLP. By contrast, both BT and FGD have demonstrated better consistency and reliability as more effective metrics for assessing SLP.

II. RELATED WORKS

A. Generative Gloss-free Sign Language Production

Saunders et al. [24] pioneered the application of Progressive Transformers (PT) to gloss-free SLP. Their approach

[†] Corresponding author.

979-8-3503-9494-8/24/$31.00 ©2024 IEEE

combined a counter decoding method and augmentation strategies such as Gaussian Noise and Future Prediction. Additionally, they introduced BT to assess the model performance. Building on this, Saunders et al. [23] addressed the regression-to-the-mean issue by adopting an adversarial training framework. In their subsequent work, they optimized PT by employing a mixture of motion primitives [25]. Meanwhile, Hwang et al. [13] introduced a paradigm shift with their Non-Autoregressive Sign Language Production with a Gaussian space (NSLP-G) model. Designed to convert spoken language sentences into corresponding sign pose sequences, NSLP-G diverged from conventional methods by adopting non-autoregressive decoding with the pre-trained VAE on the spatial aspect of the sign pose sequences. In our work, we extend this exploration into the spatial-temporal aspect of sign pose sequences, aiming to achieve a gloss-level representation.

B. Discrete Representation

There are several studies that convert continuous data into discrete data. Maddison et al. [20] and Jang et al. [14] propose Concrete Distribution and Gumbel Softmax Relaxation, respectively, which are techniques for approximating the sampling process of discrete data from a continuous distribution using annealing during training. Van et al. [29] propose VQ-VAE, which extends the standard autoencoder by adding a discrete codebook component to the network. VQ-VAE compares the vector in the codebook with the output of the encoder, where the closest vector is fed to the decoder. The model is trained using an online cluster assignment procedure coupled with a straight-through estimator. Gumbel Softmax Relaxation allows the model to effectively learn a discrete latent distribution [21]. In our work, we utilize this method to discretize the sign pose sequences.

III. METHOD

A. Problem Formulation

Consider a spoken language sentence $x = \{x_u\}_{u=1}^{U}$, which consists of U words. The objective of SLP is to produce a sign pose sequence $y = \{y_t\}_{t=1}^{T} \in \mathbb{R}^{V \times C}$, where V denotes the number of vertices, and C the feature dimension of the skeletal pose data. Instead of directly modeling $p(y|x)$, we employ an intermediary representation z, consisting of discrete tokens. These tokens encapsulate both spatial and temporal attributes of sign language. The generation process is then defined by the joint distribution $p(y,z|x) = q(y|z,x)p(z|x)$. Here, $p(z|x)$ denotes the probability of generating the discrete representation z from the input x, while $q(y|z,x)$ represents the probability of generating the continuous sign pose sequence y based on z and x. The first term is handled by a vector quantization model, and the second is modeled in an autoregressive manner.

B. Learning Discrete Representation of Sign Poses

To convert a sign pose sequence into discrete tokens, we introduce Spatio-Temporal Graph Pyramid (STGP)-based dVAE [21], as shown in Fig. 1. Inspired by the Graph Pyramid [30], we propose an STGP block–a fundamental building block for the encoder and decoder–to address the intricacies of down-sampling and up-sampling within a skeleton graph, which features non-uniform grids. A detailed overview of the STGP block is provided in Fig. 2. The STGP block sequentially processes input sequences, handling both their spatial and temporal aspects. Each processing ends with a residual connection to preserve down-sampled spatio-temporal information.

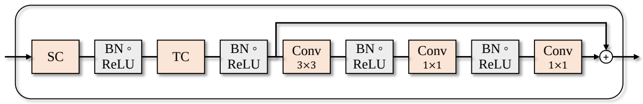


Fig. 2: An overview of the STGP block. Spatial Convolution (SC) and Temporal Convolution (TC) process the input spatially and temporally, respectively, and the subsampled output preserves spatio-temporal features through a residual connection. BN refers to Batch Normalization.

Our model is designed to handle fixed-length sign segments represented as $y^{(\ell)} \in \mathbb{R}^{L \times V \times C}$, where L represents the window size. A sign segment refers to a small fraction of the full sign pose sequence. Central to the STGP-dVAE is a codebook comprising latent variable categories, represented as $e_i \in \mathbb{R}^{K \times D_c}$. Here, K denotes the number of latent variable categories, while D_c refers to the embedding size. Then we discretize the output of the STGP encoder using Gumbel-Softmax relaxation [14]. This process samples a latent from the output of the encoder e_i as:

$$w_i = \frac{\exp(\log(e_i) + g_i)/\tau}{\sum_{k=1}^{K} \exp(\log(e_k) + g_k)/\tau}, \tag{1}$$

where g_i represents the independent i^{th} sample from the Gumbel distribution. The parameter τ adjusts the approximation to the categorical distribution, and w_i denotes the weights over the codebook vectors. These discretized representations are used for "sign tokens" in subsequent training shown in Sec. III-C. The resulting sampled latent vector is then given by $z^{(\ell)} = \sum_{k=1}^{K} w_k e_k$.

The model is optimized by minimizing the combined loss function that consists of reconstruction and diversity losses [2]. We use L2 loss as the reconstruction loss, defined as:

$$\mathscr{L}_{pose} = \frac{1}{L} \sum_{i=1}^{L} \left\| y_i^{(\ell)} - \hat{y}_i^{(\ell)} \right\|_2^2, \tag{2}$$

where $y^{(\ell)}$ and $\hat{y}^{(\ell)}$ represent the ground-truth and the reconstructed sign segment, respectively. The diversity loss, which enables the model to use the codebook effectively [2], is represented as:

$$\mathscr{L}_{div} = -\sum_{i=1}^{K} p(e_i) \log\left(p(e_i)\right). \tag{3}$$

The final loss can be defined as:

$$\mathscr{L} = \mathscr{L}_{pose} + \alpha \mathscr{L}_{div}, \tag{4}$$

where α is the hyperparameter that determines the scale of diversity loss.

TABLE I: Statistics of Sign Language Datasets. NoF refers to the number of frames.

	Train / Valid / Test	Max NoF	Min NoF	Avg NoF	FPS
PHOENIX14T [3]	7,096 / 519 / 642	475	16	116	25
How2Sign [10]	31,128 / 1,741 / 2,322	2,579	32	173	30

C. Autoregressive Sign Language Production

To generate the sign tokens from the given spoken language sentence, we use the Transformer encoder-decoder architecture, as depicted in Fig. 1. The first step involves converting the sign pose sequence into the sign tokens. This process entails dividing the input sign pose sequence, y, into multiple sign segments, each with a length L. Consequently, the number of sign tokens can be $M = \lfloor \frac{T}{L} \rfloor$, resulting in $y \approx \{y_i^{(\ell)}\}_{i=1}^{M} \in \mathbb{R}^{M \times V \times C}$. For computational efficiency, any remaining sign poses are simply removed. Each segment is subsequently encoded by the pre-trained STGP encoder through the argmax operation, yielding a sign token denoted as $z_i = \text{argmax}(h_i)$, where h_i is the i^{th} hidden representation from the STGP encoder. To mark the start and end of signs, z is padded with $\langle \text{bos} \rangle$ and $\langle \text{eos} \rangle$.

The model is optimized by minimizing the combined loss function, which consists of Cross-Entropy (CE) and latent alignment losses. The CE loss is represented as:

$$\mathscr{L}_{ce} = -\sum_{i=1}^{M} \log(p(z_i | z_{<i}, x)), \tag{5}$$

where $p(z_i|z_{<i}, x)$ represents the probability of generating the i^{th} token z_i given the previous tokens $z_{<i}$ and the input x.

The latent loss, employing L2 loss, offers supplementary latent-level signals by aligning the output of the Transformer decoder with that of the pre-trained STGP encoder. It can be defined as:

$$\mathscr{L}_{latent} = \frac{1}{M} \sum_{i=1}^{M} \left|\left| h_i - \hat{h}_i \right|\right|_2^2, \tag{6}$$

where $\hat{h}_i$ is the i^{th} hidden representation from the Transformer decoder. The overall loss is the sum of the two aforementioned losses:

$$\mathscr{L} = \mathscr{L}_{ce} + \beta \mathscr{L}_{latent}, \tag{7}$$

where β serves as a hyperparameter scaling the latent loss.

IV. EXPERIMENTAL SETTINGS

A. Datasets

We evaluated our method using two different sign language datasets: RWTH-PHOENIX-WEATHER-2014T (PHOENIX14T) [3] and How2Sign [10]. Details for each dataset are presented in Tab. I. PHOENIX14T is a German Sign Language (DGS) dataset from weather forecasts. This dataset contains 8,257 pairs of German and corresponding DGS videos with word-level annotations. How2Sign [10] is a large-scale American Sign Language (ASL) dataset that contains 2,500 instructional videos. For PHOENIX14T, where keypoints are not provided, we used OpenPose [5] and skeleton correction model [34], following [13], [24].

TABLE II: Comparison with the previous methods. The best results are in bold, followed by the second-best in underline.

	PHOENIX14T			How2Sign		
	FGD↓	DTW-MJE↓	BLEU-4	FGD↓	DTW-MJE↓	BLEU-4
PT [24]	360.62	0.793	0.59	391.06	0.355	0.57
w/o GN&FP	384.23	0.991	0.73	383.20	0.402	0.33
NSLP-G [13]	150.28	**0.638**	5.56	291.62	0.327	0.41
w/o Finetuning	179.40	0.646	4.41	440.95	0.321	0.54
SignVQNet	**92.64**	0.671	**6.85**	82.76	0.319	**0.71**
w/o Beam search	96.60	0.670	6.39	**81.99**	**0.317**	0.63
Ground-truth	0.0	0.0	8.10	0.0	0.0	0.70

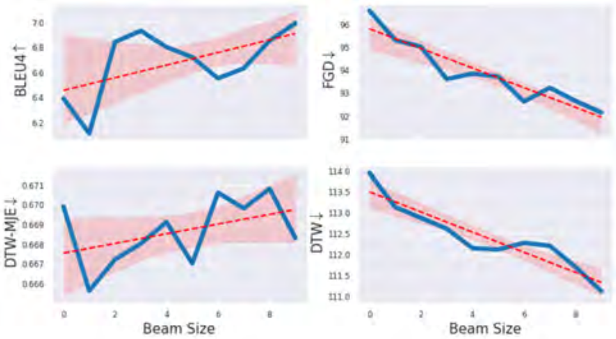


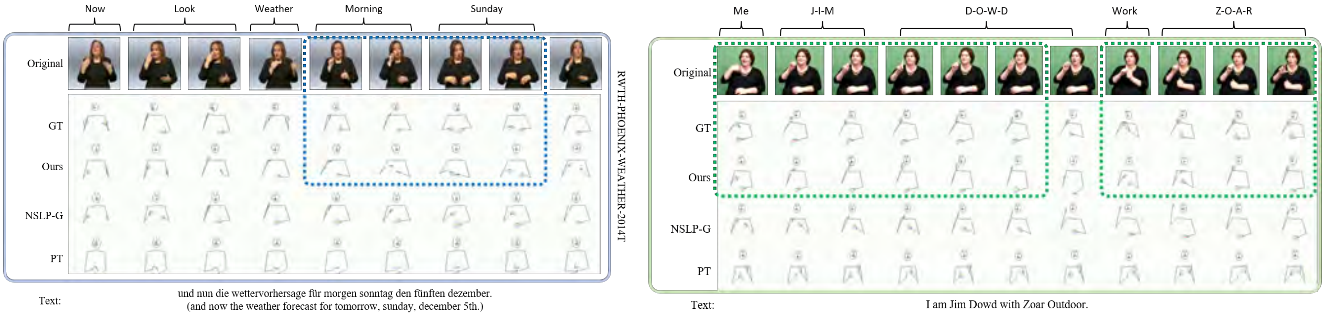
Fig. 3: Performance discrepancy among the SLP metrics in relation to change in beam size. Except for DTW-MJE, all metrics show consistent improvement as beam size increases.

B. Preprocessing

During preprocessing, to ensure consistency across all poses, the keypoints were centered and normalized relative to the shoulder joint [33]. This step guarantees that the length from one shoulder to the other is consistently scaled to a value of 1. To further refine the quality of the data, we also implemented a noise frame removal process. This process starts by calculating the differences between consecutive frames in the keypoints, represented as $X_{\text{diff}} \in \mathbb{R}^{(T-1) \times V \times C}$. Here, T denotes the number of frames, V the number of vertices, and C the feature dimension. Subsequently, we calculated the Euclidean distance for each joint between consecutive frames, resulting in a distance matrix $D \in \mathbb{R}^{(T-1) \times V}$. We then computed the average distance per frame $\overline{D}$, and compared this against a predefined threshold θ. Frames where $\overline{D}$ exceeds θ are identified as noisy and excluded from further processing. Following the removal of noisy frames, we normalized the remaining keypoints to ensure that they fit within the range of $[-1, 1]$. This final normalization step is essential for maintaining uniform scaling and positioning of the keypoints. Texts were converted to lowercase and tokenized via Byte-Pare Encoding (BPE). The vocabulary sizes for this encoding were set at 3,000 for PHOENIX14T dataset and 10,000 for How2Sign dataset.

C. Implementation Details

For our experiments, we utilized the Gumbel-Softmax relaxation and annealed its temperature, τ, from 0.9 down to 0.1. The parameters α and β were set at 0.1 and 0.001, respectively. We used 4 STGP blocks. In configuring the Transformer model, we set the hidden size to 768, with 4 layers and 8 attention heads, a dropout rate to 0.1, and an intermediate size to 1,024. We used the AdamW optimizer


Now
Look
Weather
Morning
Sunday
Original
GT
Ours
NSLP-G
PT
RWTH-PHOENIX-WEATHER-2014T
Text:
und nun die wettervorhersage für morgen sonntag den fünften dezember.
(and now the weather forecast for tomorrow, sunday, december 5th.)
Me
J-I-M
D-O-W-D
Work
Z-O-A-R
Original
GT
Ours
NSLP-G
PT
How2Sign
Text:
I am Jim Dowd with Zoar Outdoor.

(a) PHOENIX14T

(b) How2Sign

Fig. 4: We present a visual comparison between our method and the baselines on both (a) PHOENIX-2014-T and (b) How2Sign. As highlighted in the dashed boxes, our method generates more realistic and accurate sign pose sequences. Videos are available at http://nlpcl.kaist.ac.kr/ projects/signvqnet

TABLE III: Ablation experiments on PHOENIX14T.

Window Size	DEV FGD↓	 DTW-MJE↓	 BLEU-4	TEST FGD↓	 DTW-MJE↓	 BLEU-4
16	42.95	0.310	6.75	41.45	0.301	6.55
32	**42.04**	**0.302**	**6.77**	**41.43**	**0.299**	**6.88**
64	44.68	0.305	6.44	43.21	0.305	6.52

(a) Window Size

Loss	DEV FGD↓	 DTW-MJE↓	 BLEU-4	TEST FGD↓	 DTW-MJE↓	 BLEU-4
$\mathscr{L}_{ce}$	102.87	0.679	6.30	**96.51**	0.674	5.94
$\mathscr{L}_{latent}$	504.88	0.755	0.36	512.48	0.741	0.40
$\mathscr{L}_{ce} + \mathscr{L}_{latent}$	**100.85**	**0.676**	**6.76**	96.60	**0.670**	**6.39**

(b) Loss Type

Models	DEV FGD↓	 DTW-MJE↓	 BLEU-4	TEST FGD↓	 DTW-MJE↓	 BLEU-4
GRU	433.76	0.685	0.49	444.74	0.681	0.61
+Attn	383.99	0.697	0.74	402.92	0.700	0.81
Transformer	103.88	0.682	5.33	102.45	0.676	**6.22**
+BERT [8]	**102.87**	**0.679**	**6.30**	**96.51**	**0.674**	5.94

(c) Architecture Type

Vocab Size	DEV FGD↓	 DTW-MJE↓	 BLEU-4	TEST FGD↓	 DTW-MJE↓	 BLEU-4
512	42.96	0.315	6.75	41.81	0.312	6.30
1024	**42.04**	**0.302**	**6.77**	**41.43**	**0.299**	**6.88**
2048	44.68	0.306	6.24	44.23	0.306	6.77
4096	56.74	0.325	6.64	55.13	0.323	6.57
8192	55.96	0.317	6.30	55.58	0.315	5.89

(d) Codebook Size

[19] with a learning rate set at 0.0001. To encode the spoken language setnecne, we used the pre-trained BERT[1] (bert-base-cased and bert-base-german-cased), fine-tuned during training. We selected a checkpoint that minimizes the score against FGD metric. The entire training process ran for 500 epochs, taking approximately 24 hours on a Tesla A100 GPU, with a batch size of 64.

D. Evaluation Metrics

We used a range of evaluation metrics to assess our method. The Back-Translation (BT) [24] was used to compute BLEU-4 by translating the produced sign pose sequence back into spoken language for comparison with the original text. As a back-translation model, we trained Joint-SLT [4] on PHOENIX14T and How2Sign, following [12], [13], [24], [28]. In addition, we used the DTW-MJE [12], which combines Dynamic Time Warping (DTW) with Mean Joint Error (MJE) to measure discrepancies between predicted and actual sign pose sequences. Additionally, we used Fréchet Gesture Distance (FGD) [32] to evaluate the visual fidelity of the generated sign pose sequence by comparing the distributions of real and generated sequences.

V. Experimental Results

A. Quantitative Results

We compared our method with the previous gloss-free SLP methods: PT [24] and NSLP-G [13]. For PT, we employ its base and Gaussian and Future Prediction (GN&FP) settings. PT was modified to exclude the use of additional ground-truth data, as recommended by [12], [28], to ensure a fair comparison. For NSLP-G, we used the frozen and fine-tuning option during training. As shown in Tab. II, our method outperforms the baselines on both datasets, especially in terms of FGD and BLEU-4. On How2Sign, which features longer frame sequences than PHOENIX14T (Tab. I), our method shows its robustness in generating extended pose sequences. An interesting observation arises from PHOENIX14T. While our method outperforms in FGD and BLEU-4, it lags behind NSLP-G in DTW-MJE. This is mainly due to the unique evaluating manner in DTW-MJE. We delve deeper into this observation in the subsequent sections.

B. Analyzing Evaluation Metrics for SLP

In the domain of SLP, selecting reliable metrics is crucial for the comprehensive evaluation of the performance of generative models. While the field currently relies on metrics such as FGD, DTW-MJE, and BT, identifying the optimal metric for comprehensive evaluation remains an open question. This challenge is exemplified by the conflicting results observed between our method and NSLP-G, as highlighted in Sec. V-A and by [1]. Specifically, these discrepancies arise from the differences in loss functions employed by these models. For instance, our method uses CE loss, which focuses on sequential prediction accuracy and preserving the structure of sign language. This emphasis on the sequential

[1] https://huggingface.co/models

structure may affect the model's performance in DTW-MJE, a metric that primarily evaluates the spatial accuracy of keypoints. By contrast, NSLP-G utilizes MSE loss, focusing on the spatial accuracy on a frame-by-frame basis, which typically results in better scores in DTW-MJE.

To further investigate these disparities, we conducted an additional analysis to see how varying beam sizes affects the performance of our model on PHOENIX14T. Our findings, as shown in Fig. 3, indicate that FGD and BLEU-4 generally improve with larger beam sizes, whereas DTW-MJE tends to decrease. Additionally, we included DTW in our analysis for a more comprehensive evaluation. It is worth noting that DTW-MJE, by enforcing alignment between the generated and ground-truth sign pose sequences, might not always provide a true comparison. Specifically, DTW aims to minimize the distance by aligning sequences, which may not always reflect the actual temporal alignment. When combined with MJE, this can lead to inconsistent error measurements. This shows the need for careful consideration of the development of new metrics tailored to the specific challenges of SLP evaluation.

C. Qualitative Results

We offer a visual comparison of the sign pose sequences generated by our method in contrast to the baselines on PHOENIX14T and How2Sign. As shown in Fig. 4, our method generates more accurate signs, such as "Morning", "Sunday", "Me", and "Work", compared to the baselines.

D. Ablation Study

We investigated the effects of various components and design choices of our method on PHOENIX14T, the most widely used dataset in sign language research. Tab. IIIc shows that the Transformer encoder-decoder model outperforms GRU-based networks, which are commonly used for human motion tasks [11], [22], [32]. Furthermore, employing the pre-trained BERT [8] significantly improved its performance. Regarding the loss functions, a combination of $\mathscr{L}_{ce}$ and $\mathscr{L}_{latent}$ yields the best performance, as shown in Tab. IIIb. Optimal performance was achieved with a window size L set to 32 as shown in Tab. IIIa, and a codebook size of 1,024 showed the best performance as shown in Tab. IIId.

VI. CONCLUSION

In this paper, we introduced SignVQNet, gloss-free Sign Language Production (SLP) that leverages vector quantization to derive discrete tokens from sign pose sequences. This enables genuine autoregression without the need for ground-truth data during inference, addressing the shortcomings of the previous autoregressive SLP model. In our experiments, we demonstrate its superiority over prior methods, achieving the state-of-the-art performance on both PHOENIX14T and How2Sign. Additionally, we highlight the reliability of BT and FGD as evaluation metrics, while noting inconsistencies in the DTW-MJE metric.

VII. ACKNOWLEDGMENTS

This work was supported by the Technology Innovation Program (20014406, Development of interactive sign language interpretation service based on artificial intelligence for the hearing impaired) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea).

REFERENCES

[1] R. S. Arkushin, A. Moryossef, and O. Fried. Ham2pose: Animating sign language notation into pose sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21046–21056, 2023.

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

[3] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7784–7793, 2018.

[4] N. C. Camgöz, O. Koller, S. Hadfield, and R. Bowden. Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10023–10033, 2020.

[5] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. OpenPose: Realtime Multi-person 2D Pose Estimation using Part Affinity Fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019.

[6] Y. Cheng, F. Wei, J. Bao, D. Chen, and W. Zhang. Cico: Domain-aware sign language retrieval via cross-lingual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19016–19026, 2023.

[7] S. Cho, S. Jeong, J. y. Seo, and J. Park. Discrete prompt optimization via constrained generation for zero-shot re-ranker. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 960–971, Toronto, Canada, July 2023. Association for Computational Linguistics.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[9] A. Duarte, S. Albanie, X. Giró-i Nieto, and G. Varol. Sign language video retrieval with free-form textual queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14094–14104, 2022.

[10] A. Duarte, S. Palaskar, L. Ventura, D. Ghadiyaram, K. DeHaan, F. Metze, J. Torres, and X. Giro-i Nieto. How2sign: A large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2735–2744, 2021.

[11] A. Gopalakrishnan, A. Mali, D. Kifer, L. Giles, and A. G. Ororbia. A neural temporal model for human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12116–12125, 2019.

[12] W. Huang, W. Pan, Z. Zhao, and Q. Tian. Towards fast and high-quality sign language production. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3172–3181, 2021.

[13] E. J. Hwang, J.-H. Kim, and J. C. Park. Non-autoregressive sign language production with gaussian space. In *Proceedings of the 32nd British Machine Vision Conference*, pages 22–25, 2021.

[14] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[15] S. Jeong, J. Baek, C. Park, and J. Park. Unsupervised document expansion for information retrieval with stochastic text generation. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 7–17, Online, June 2021. Association for Computational Linguistics.

[16] H. Lee, J.-H. Kim, E. J. Hwang, J. Kim, and J. C. Park. Leveraging large language models with vocabulary sharing for sign language translation. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5, 2023.

[17] D. Li, C. Xu, X. Yu, K. Zhang, B. Swift, H. Suominen, and H. Li. Tsp-net: Hierarchical feature learning via temporal semantic pyramid for sign language translation. *Advances in Neural Information Processing Systems*, 33:12034–12045, 2020.

[18] K. Lin, X. Wang, L. Zhu, K. Sun, B. Zhang, and Y. Yang. Gloss-free end-to-end sign language translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 12904–12916, 2023.

[19] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017.

[20] C. J. Maddison, A. Mnih, and Y. W. Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

[21] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831, 2021.

[22] H.-F. Sang, Z.-Z. Chen, and D.-K. He. Human motion prediction based on attention mechanism. *Multimedia Tools and Applications*, 79(9):5529–5544, 2020.

[23] B. Saunders, N. C. Camgoz, and R. Bowden. Adversarial training for multi-channel sign language production. In *Proceedings of the 31st British Machine Vision Conference*, pages 7–10, 2020.

[24] B. Saunders, N. C. Camgoz, and R. Bowden. Progressive transformers for end-to-end sign language production. In *Proceedings of the 16th European Conference on Computer Vision*, pages 687–705. Springer, 2020.

[25] B. Saunders, N. C. Camgoz, and R. Bowden. Mixed signals: Sign language production via a mixture of motion primitives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1919–1929, 2021.

[26] B. Saunders, N. C. Camgoz, and R. Bowden. 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*, pages 5141–5151, 2022.

[27] B. Shi, D. Brentari, G. Shakhnarovich, and K. Livescu. Open-domain sign language translation learned from online video. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6365–6379, 2022.

[28] S. Tang, R. Hong, D. Guo, and M. Wang. Gloss semantic-enhanced network with online back-translation for sign language production. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5630–5638, 2022.

[29] A. Van Den Oord, O. Vinyals, et al. Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 30, 2017.

[30] S. Yan, Z. Li, Y. Xiong, H. Yan, and D. Lin. Convolutional sequence generation for skeleton-based action synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4394–4402, 2019.

[31] A. Yin, T. Zhong, L. Tang, W. Jin, T. Jin, and Z. Zhao. Gloss attention for gloss-free sign language translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2551–2562, 2023.

[32] Y. Yoon, B. Cha, J.-H. Lee, M. Jang, J. Lee, J. Kim, and G. Lee. Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity. *ACM Transactions on Graphics (TOG)*, 39(6):1–16, 2020.

[33] Y. Yoon, W.-R. Ko, M. Jang, J. Lee, J. Kim, and G. Lee. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4303–4309. IEEE, 2019.

[34] J. Zelinka and J. Kanis. Neural Sign Language Synthesis: Words are our Glosses. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3395–3403, 2020.