Gloss-driven Conditional Diffusion Models for Sign Language Production

SHENGENG TANG, School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China
FENG XUE, School of Software, Hefei University of Technology, Hefei, China
JINGJING WU, School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China
SHUO WANG, School of Data Science, School of Information Science and Technology, University of Science and Technology of China, Hefei, China
RICHANG HONG, School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China

Sign Language Production (SLP) aims to convert text or audio sentences into sign language videos corresponding to their semantics, which is challenging due to the diversity and complexity of sign languages, and cross-modal semantic mapping issues. In this work, we propose a Gloss-driven Conditional Diffusion Model (GCDM) for SLP. The core of the GCDM is a diffusion model architecture, in which the sign gloss sequence is encoded by a Transformer-based encoder and input into the diffusion model as a semantic prior condition. In the process of sign pose generation, the textual semantic priors carried in the encoded gloss features are integrated into the embedded Gaussian noise via cross-attention. Subsequently, the model converts the fused features into sign language pose sequences through T-round denoising steps. During the training process, the model uses the ground-truth labels of sign poses as the starting point, generates Gaussian noise through T rounds of noise, and then performs T rounds of denoising to approximate the real sign language gestures. The entire process is constrained by the MAE loss function to ensure that the generated sign language gestures are as close as possible to the real labels. In the inference phase, the model directly randomly samples a set of Gaussian noise, generates multiple sign language gesture sequence hypotheses under the guidance of the gloss sequence, and outputs a high-confidence sign language gesture video by averaging multiple hypotheses. Experimental results on the Phoenix2014T dataset show that the proposed GCDM method achieves competitiveness in both quantitative performance and qualitative visualization.

CCS Concepts: • **Computing methodologies** → **Computer vision**; **Natural language processing**; *Machine learning*.

Additional Key Words and Phrases: Sign Language Production, Gloss Semantic Encoding, Diffusion Model, Deep Learning

*Corresponding authors.

Authors' addresses: S. Tang, F. Xue (Corresponding author), J. Wu, R. Hong (Corresponding author), Hefei University of Technology, No. 485 Danxia Road, Hefei, Anhui, China, 230601, e-mails: tangsg@hfut.edu.cn, feng.xue@hfut.edu.cn, hfutwujingjing@mail.hfut.edu.cn, hongrc@hfut.edu.cn; S. Wang, University of Science and Technology of China, No. 96, JinZhai Road, Hefei, Anhui, China, 230026, email: shuowang.edu@gmail.com.

ACM 1551-6865/2024/5-ART
https://doi.org/10.1145/3663572

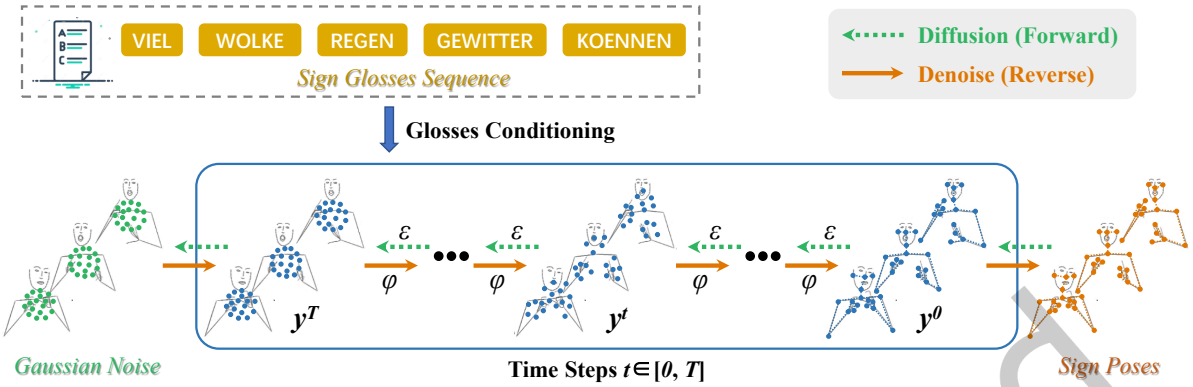

Fig. 1. The basic idea of the GCDM for SLP. The proposed GCDM takes the sign gloss sequence encoding as a condition and reversely diffuses Gaussian distribution sampling by T-step denoising into sign poses.

1 INTRODUCTION

Sign Language Production (SLP) is a new emerging and challenging task in the computer vision-language field, related to natural language processing [69, 73], human pose analysis [2, 12, 30], video analysis [34, 59, 60] and cross-media reasoning [33, 47, 48], etc. Specifically, SLP is the inverse process of Sign Language Recognition (SLR) [15, 19, 52], which converts textual sentences into visual representations of sign language. This task requires the model to understand textual semantics and generate matching sign representations (*i.e.*, sign pose or video) based on the semantics.

Sign glosses are spoken language words matching the meaning of signs, which are defined as minimal lexical items. As the basic semantic unit in sign language, gloss plays a crucial transitional role in SLP. Previous SLP works usually first translate the spoken language into a gloss sequence through the machine translation model, and then convert the glosses into a series of sign poses (G2P) [42, 43, 49]. Since G2P is a cross-media task involving both textual understanding and visual generation, it is more challenging and decisive for the success of SLP. In this work, we focus on G2P, the core procedure of the SLP task.

For SLP, early works [24, 25] mainly utilize avatar-based and Statistical Machine Translation (SMT) methods, which require expensive pose pre-capture and struggle to cope with non-matching phrases. Recent efforts towards SLP try to model the text-to-vision mapping in sign language using deep neural networks [31, 63, 68]. Given the excellent performance of Generative Adversarial Networks (GANs) on generative tasks, some SLP methods based on conditional GANs have emerged [42, 45]. These methods generate sign language representations from textual inputs and optimize SLP by discriminating the authenticity of sequences (*i.e.*, original or generated). Additionally, some work is devoted to exploring Non-AutoRegressive models to address high inference latency and error propagation problems in SLP [21, 22]. Another common practice is to use a Transformer-based encoding-decoding framework that first encodes the textual inputs and then decodes the textual embeddings into pose sequences of a given length [43, 67]. The above methods focus on straightforwardly generating visual representations of sign language from text sequences, which ignores the complexity of text-to-visual cross-modal conversion. In fact, generating complex dynamic gestures based on scarce textual semantic cues should be a step-by-step process. It is necessary to gradually approximate the target gesture under semantic guidance, which makes sign language generation more flexible and finely controlled.

To this end, we propose a novel Gloss-driven Conditional Diffusion Model (GCDM) for G2P in the SLP task. As shown in Figure 1, the core of the proposed GCDM is a diffusion model architecture, in which the sign gloss

sequence is encoded by a Transformer-based encoder and input into the diffusion model as a semantic prior condition. In the process of sign pose generation, the textual semantic priors carried in the encoded gloss features are integrated into the embedded Gaussian noise via cross-attention. Subsequently, the model converts the fused features into sign language pose sequences through T-round denoising steps. Besides, a multi-hypothesis aggregation mechanism is introduced during the inference phase to generate the higher-confidence sign language pose video. Our main contributions can be summarized as follows:

- We propose a novel Gloss-driven Conditional Diffusion Model (GCDM) for SLP, which gradually removes noise in Gaussian distribution samples to obtain sign pose videos, driven by the gloss semantic prior condition.
- A multi-hypothesis aggregation mechanism is introduced in the inference phase, which generates multiple sign language video hypotheses under the guidance of the gloss condition, and outputs higher-confidence sign poses by averaging the above hypotheses.
- Extensive experiments on the challenging PHOENIX14T [3] dataset demonstrate the superiority of the proposed method. Ablation studies and qualitative visualizations also verify the contribution of each component.

2 RELATED WORK

2.1 Sign Language Production

Over the past four decades, sign language research has evolved from isolated Sign Language Recognition (SLR) [9, 13, 14, 18, 61], continuous Sign Language Translation (SLT) [3, 16, 17, 66, 74], to Sign Language Production (SLP) [8, 21, 22, 43, 46, 49, 53]. Previous SLP works have focused on avatar-based [11, 24] and Statistical Machine Translation (SMT) [25, 29] methods, which can generate realistic sign gestures. However, these methods rely on rule-based lookup of phrases in pre-captured motion databases, thus requiring expensive preprocessing costs and limited by predefined phrases.

Recently, an increasing number of deep learning models have been applied to SLP tasks, such as RNN-based models [8, 67, 68], Generative Adversarial Network (GAN) [31, 49, 50, 55], Variational Auto Encoder (VAE) [22, 63] and Transformers [21, 37, 42, 43, 45, 46, 53]. Early work on deep learning-based SLP considers directly translating textual descriptions into photo-realistic sign language video (TG2V), which struggles to handle both gesture accuracy and finger details [8, 70]. The classic approach to the solving of SLP problem is to divide it into three main sub-tasks [49, 50], namely Text-to-Gloss translation (T2G), Gloss-to-Pose generation (G2P), and Pose-to-Video synthesis (P2V). P2V is usually regarded as a pure Computer Vision (CV) problem, solved by pose-guided video synthesis techniques [38, 44, 57]. More works have focused on the G2P sub-task of SLP, which is a challenging text-to-vision cross-media task [21, 22, 42, 43, 46, 67]. For G2P, Saunders *et al.* propose a mixture of motion primitives network, which produces an infinite number of unique sign poses based on a Mixture-of-Expert (MoE) architecture [46]. To avoid the error accumulation caused by AutoRegression, Hwang *et al.* first build a Gaussian space to learn the generation of each sign pose, and then adopt a non-AutoRegressive model to map from the source sentence to the target distribution [21]. Besides, Huang *et al.* propose an external aligner based on monotonic alignment search for gloss duration prediction, and devise a spatial-temporal graph convolutional pose generator to produce smoother and more natural sign pose sequences [21].

The Transformer [56] is a sequence-to-sequence learning model based solely on the attention mechanism, which can transform source sequences into target representations with global dependencies. Transformer-based models are originally used in the field of Natural Language Processing (NLP), especially in Neural Machine Translation (NMT) [39, 54, 64]. Due to the great success of the transformer in NLP, researchers have tried to generalize it to a wider field. Some classic computer vision and vision-language models based on Transformer came into being, such as ViT [10], ViViT [1], DERT [5], Deformable DETR [75], VL-BERT [51], and CLIP [40]. As a

typical sequence modeling task, SLP shares similar nature with the abovementioned tasks, thus Transformer-based models have been widely used for SLP [21, 42, 43, 45, 46, 58, 67]. Saunders *et al.* design a progressive transformer to generate sign poses in an end-to-end manner [43]. Going a step further, they introduce the adversarial training scheme into the Transformer framework, learning to distinguish between real and fake sequences to ensure the production of realistic and expressive poses [42]. Zelinka *et al.* devise the feed-forward transformer and recurrent transformer to convert the input Czech text into a sequence of skeletal poses [67]. In addition, Viegas *et al.* propose a dual encoder Transformer able to generate manual signs as well as facial expressions from both sign text and gloss annotations [58].

However, the above methods focus on tackling the conversion from glosses to poses and fitting between output poses and ground truth, while ignoring the mining of semantic cues from the textual input, which may make it difficult for the obtained poses to cover the original intended meaning. In contrast, our approach is devoted to strengthening semantic learning of the source sentence and enhancing the traction role of textual clues during the pose generation.

2.2 Diffusion Model

Diffusion models have recently emerged as a powerful class of generative models, offering a promising alternative to traditional approaches like GANs [31, 50, 55] and VAEs [22, 63]. These models, inspired by the physics of diffusion processes, aim to learn the gradual transition from a simple, unstructured distribution to a complex, data-like distribution. The key idea behind diffusion models is to reverse a diffusion process that gradually adds noise to the data, thus generating novel and realistic samples. Diffusion models originate from applications in the field of computer vision generation and have also achieved great success in tasks such as natural language generation [7, 72], multi-modal learning [26, 65], and waveform signal processing [6, 28].

For some time, diffusion models have demonstrated remarkable capabilities, especially in the domain of image synthesis [41, 71]. However, their application to video generation, particularly for complex tasks like human pose video generation, remains relatively unexplored. Human pose video generation presents unique challenges, such as maintaining temporal coherence and generating realistic movements across frames. Several studies have attempted to tackle these challenges using various techniques. For instance, Hasegawa *et al.* [20] propose a method that combines convolutional neural networks with recurrent neural networks to generate sequences of human poses. While their approach achieves some degree of temporal coherence, it often struggles with generating realistic movements, especially for long sequences. More recently, Luo *et al.* [36] introduce a diffusion-based model specifically designed for video generation. Their model, while showing promising results in general video synthesis tasks, does not directly address the specific challenges of human pose generation in a targeted manner.

This work builds upon these prior efforts by introducing a novel diffusion model tailored for sign language pose video generation. Our approach incorporates gloss semantic priors and leverages the expressive power of diffusion models to generate realistic and coherent sign pose sequences. By gradually refining the modeling of sign language poses, our model achieves superior performance compared to previous methods.

2.3 Multi-Hypothesis Aggregation

In the realm of sign language pose generation for video, multi-hypothesis aggregation plays a crucial role in synthesizing accurate and consistent pose sequences. While advanced techniques have been explored, a surprisingly common approach still involves straightforward averaging or taking the optimal solution of multiple pose hypotheses. These approaches have been widely used in previous works, such as [32, 62]. The reason for its popularity lies in its computational efficiency and ease of implementation. However, as the field of sign language pose generation evolves, so must the aggregation strategies.

Table 1. Notations and Definitions in Our Model.

Symbol	Description
$X = \{x_1, x_2, \cdots, x_N\}$	Input textual sentence of the gloss encoder with N glosses
$Y = \{y_1, y_2, \cdots, y_U\}$	Target sign pose sequence with U frames (*i.e.*, ground truth)
$\widetilde{G} = \{\tilde{g}_1, \tilde{g}_2, \cdots, \tilde{g}_N\}$	Gloss embeddings with positional encoding
$\widetilde{X} = \{\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_N\}$	Gloss encodings output from the gloss encoder (*i.e.*, gloss condition)
$\widehat{Y} = \{\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_U\}$	Diffused poses with noise / Noises sampled from Gaussian distribution
$\widetilde{H} = \{\tilde{h}_1, \tilde{h}_2, \cdots, \tilde{h}_U\}$	Noisy pose stream embedded with gloss condition in pose denoiser
$\widetilde{Y} = \{\tilde{y}_1, \tilde{y}_2, \cdots, \tilde{y}_U\}$	Generated sign pose sequence with U frames

More recent works have attempted to address these limitations by exploring more sophisticated aggregation techniques. Li *et al.* [35] design a cross-hypothesis interaction module to enable interactions among multi-hypothesis features, thereby aggregating the multi-hypothesis features to synthesize the final 3D pose. Saunders *et al.* [46] propose a mixture of motion primitives architecture for sign language animation, in which a set of distinct motion primitives are learned to be temporally combined at inference to animate continuous sign language sequences.

Our method, while utilizing the averaging approach for simplicity, recognizes the need for further refinement. We aim to enhance the accuracy and naturalness of the generated poses by incorporating additional considerations, such as joint-level differences and the exploitation of 2D keypoint information. Future work in this area could focus on developing more advanced aggregation strategies that combine the benefits of simplicity with the accuracy afforded by more complex techniques.

3 METHOD

Given a sign sentence $X = \{x_1, x_2, \cdots, x_N\}$ with N glosses, our SLP system aims to generate the semantically corresponding sign pose sequence $\widetilde{Y} = \{\tilde{y}_1, \tilde{y}_2, \cdots, \tilde{y}_U\}$ with U frames. Furthermore, we take ground-truth poses $Y = \{y_1, y_2, \cdots, y_U\}$ as fitting targets in SLP. To clarify the data stream in the GCDM framework, we elaborate on these notations in Table 1.

3.1 Overall Pipeline

The overall pipeline of the proposed Gloss-driven Conditional Diffusion Model (GCDM) is illustrated in Figure 2, whose execution mainly includes two phases: a training phase (forward diffusion and reverse diffusion based on target labels, see Section 3.2) and an inference phase (reverse diffusion from pure noise samples, see Section 3.3). Specifically, the proposed GCDM is based on a diffusion model architecture, in which the sign gloss sequence is encoded by a Transformer-based encoder and input into the diffusion model as a semantic prior condition. In the process of sign pose generation, the textual semantic priors carried in the encoded gloss features are integrated into the embedded Gaussian noise via cross-attention. Subsequently, the model converts the fused features into sign language pose sequences through T-round denoising steps.

During the training phase, the model uses the ground-truth labels of sign poses as the starting point, generates Gaussian noise through T rounds of noise, and then performs T rounds of denoising to approximate the real sign language poses. The entire process is constrained by the MAE loss function to ensure that the generated sign language gestures are as close as possible to the target labels. In the inference phase, the model directly randomly samples a set of Gaussian noise, generates multiple sign language gesture sequence hypotheses under

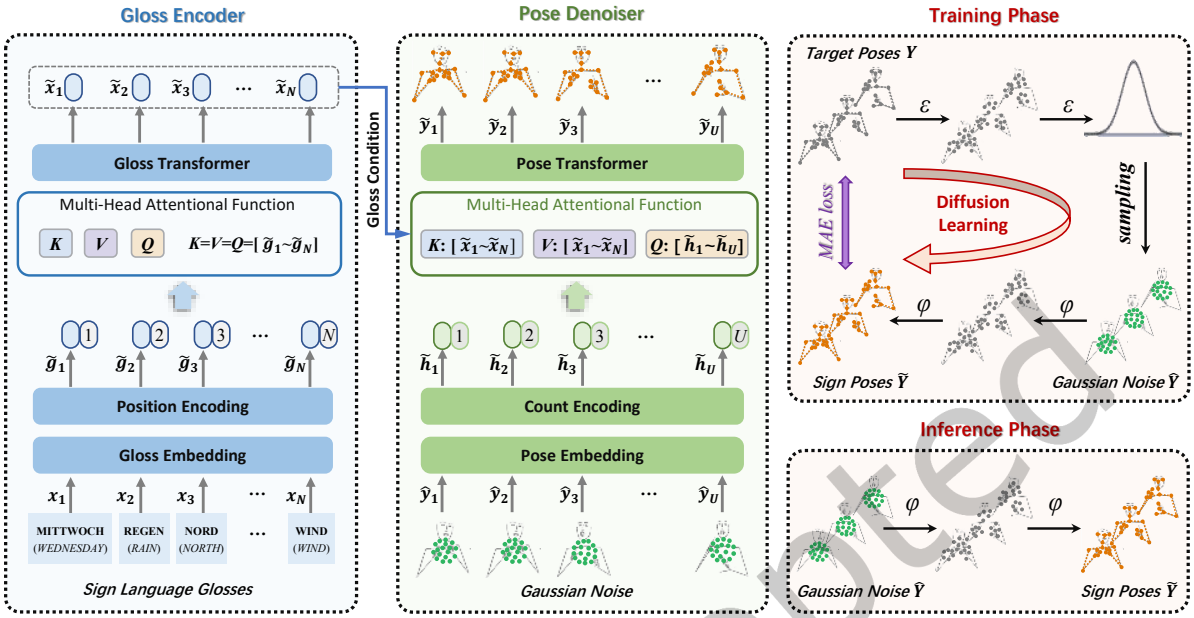

Fig. 2. Overview of the proposed GCDM. It consists of a gloss encoder and a pose denoiser. Thereinto, the Transformer-based gloss encoder is used to learn the semantics of the gloss sequence $X = \{x_{1:N}\}$. Then, the encoded gloss embedding is regarded as a semantic condition to guide the denoising process in diffusion learning. After T timesteps, based on the Gaussian noise samples $\tilde{Y} = \{\tilde{y}_{1:U}\}$, the pose denoiser outputs the sign pose sequence $Y = \{y_{1:U}\}$. For optimization, the MAE loss $\mathcal{L}_{MAE}$ is adopted to calculate the absolute error between the output sign poses and the target poses to evaluate the generation performance.

the guidance of the gloss sequence, and outputs a high-confidence sign language pose video by averaging multiple hypotheses.

3.2 Gloss-driven Conditional Diffusion

3.2.1 Gloss Condition Encoding. In this module, we build a transformer-based encoder, which encodes input glosses into gloss embedding tokens. To make glosses with similar semantics closer, we map the source tokens $X = \{x_n\}_{n=1}^{N}$ into a high-dimensional space using a linear embedding layer:

$$g_n = W^x \cdot x_n + b^x, g_n \in \mathbb{R}^{1\times d_x} \tag{1}$$

where g_n is the vector representation of the gloss tokens, W^x and b^x represent the weight and bias during gloss embedding, respectively.

Similar to grammatical spoken languages, sign language has its own unique linguistic rules. Considering that the self-attention mechanism cannot directly encode the temporal information, we apply a positional encoding layer to provide the temporal order of gloss vectors:

$$\tilde{g}_n = g_n + PE(n), \tag{2}$$

where PE is implemented by the predefined sine and cosine functions of different frequencies.

Our gloss encoder consists of K identical blocks, each of which includes a Multi-Head Attention (MHA), a Normalization Layer (NL), and a Feedforward Layer (FL). For ease of understanding, we use z_k to mark the

learned feature sequence after the k-th block. The calculation process of the gloss encoder can be expressed as:

$$\widetilde{X} = GlossEncoder(\widetilde{G}) \Leftrightarrow \begin{cases} z_0 = \widetilde{G} = \{\tilde{g}_1, \tilde{g}_2, \cdots, \tilde{g}_n\}; \\ z_k = FL(MHA(Q,K,V)|_{Q=K=V=NL(z_k)} + z_{k-1}), k \in [1,K]; \\ \widetilde{X} = NL(z_K). \end{cases} \tag{3}$$

Here, MHA plays a key role in aggregating gloabl token representations and computing gloss sequence contextual dependencies. Specifically, MHA computes scaled dot-product attention based on a general Multi-Head Attention (MHA) mechanism, which learns the relationship between queries and values from a series of matrices (querys matrix Q, values matrix V, keys matrix K).

$$Attention(Q,K,V) = softmax(\frac{QK^\top}{\sqrt{d}})V, \tag{4}$$

where $\frac{1}{\sqrt{d}}$ is the scaling factor, and $d_q = d_k$. In MHA, Q, K and V are all equal to z_k, so the output is a contextual sequence with self-attention.

MHA handles the above attention mechanism in parallel using M different mappings, which allows the model to capture complementary information from different representation sub-spaces. Then, the outputs of each head are concatenated and projected together through a linear layer.

$$\begin{cases} head_m = Attention(QW_m^Q, KW_m^K, VW_m^V); \\ MHA(Q,K,V) = [head_1, \cdots, head_M] \cdot W^O, \end{cases} \tag{5}$$

where QW_m^Q, KW_m^K, VW_m^V and W^O are the parameter matrices of weights related to inputs.

3.2.2 Label Forward Diffusion. We first sample a timestep $t \sim U(0,T)$, where T is the maximum number of timesteps. The forward diffusion process is to gradually add Gaussian noise to the label data Y_0 through an approximate posterior $q(Y_{1:T}|Y_0)$ modeled by a Markov chain, converting it into a Gaussian distribution $\mathcal{N}(0,\mathbf{I})$. Here, we predefine the true label Y as the initial noise-free data Y_T. At the t-th time step, the Markov process can be expressed as:

$$q(Y_t|Y_{t-1}) = \mathcal{N}(Y_t; \sqrt{1-\beta_t}Y_{t-1}, \beta_t\mathbf{I}), \tag{6}$$

where β_t is the cosine noise variance schedule. The marginal distribution of Y_t is given by:

$$\begin{aligned} q(Y_t|Y_0) &:= \mathcal{N}\left(Y_t; \sqrt{\bar{\alpha}_t}Y_0, (1-\bar{\alpha}_t)\mathbf{I}\right), \\ Y_t &= \sqrt{\bar{\alpha}_t}Y_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}(0,\mathbf{I}), \end{aligned} \tag{7}$$

where $\alpha_t := 1-\beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t}\alpha_s$. When the value of timestep is large enough, the distribution of $q(Y_T)$ is close to an isotropic Gaussian distribution. During the reverse diffusion process, Y_T is regarded as the initial noise sample in the pose denoiser, also noted as $\widehat{Y}$.

3.2.3 Semantic-driven Training. In this work, we adopt a Transformer-based pose denoiser driven by the gloss condition, which inputs the noisy poses $\widehat{Y}$ and outputs the generated pose sequence $\widetilde{Y} = \{\tilde{y}_1, \tilde{y}_2, \cdots, \tilde{y}_U\}$. Here, we denote the original input of the denoiser as noisy data. Factly, the pose denoiser aims to implement a reverse diffusion process on noisy poses to convert them into pure sign poses without noise. A linear layer is first adopted to map the noisy data $\tilde{y}_u \in \mathbb{R}^{1\times d'}$ into a high-dimensional space:

$$h_u = W^y \cdot \hat{y}_u + b^y, \hat{y}_u \in \mathbb{R}^{1\times d_y}, \tag{8}$$

where h_u is the vector representation of the noisy data, W^y and b^y represent the weight and bias during pose embedding, respectively. Similar to the PE function (see Equation 2), a count encoding layer CE is used to represent the position of each frame in the entire target sequence:

$$\tilde{h}_u = h_u + CE(u). \tag{9}$$

Different from the gloss encoder, we introduce MHA with an interactive attention mechanism in the pose denoiser to realize the guidance of the gloss semantic condition $\widetilde{X} = \{\tilde{x}_n\}_{n=1}^{N}$ for the pose streams $\widetilde{H} = \{\tilde{h}_u\}_{u=1}^{U}$. The calculation process of the pose denoiser can be expressed as:

$$\begin{aligned} &\widetilde{Y} = PoseDenoiser(\widetilde{X}, \widetilde{H}) \Leftrightarrow \\ &\begin{cases} z_1 = FL(MHA(Q,K,V)|_{Q=NL(\widetilde{H}),K=V=\widetilde{X}}); \\ z_k = FL(MHA(Q,K,V)|_{Q=K=V=NL(z_k)} + z_{k-1}), k \in [2,K]; \\ \widetilde{Y} = FL'(NL(z_K)). \end{cases} \end{aligned} \tag{10}$$

where MHA, NL and FL are the same as ones in the Equation 3. The adopted final linear layer FL' aims to map the denoised stream into the sign poses with 3D coordinates.

The whole encoder-denoiser diffusion model is trained using the Mean Absolute Error (MAE) loss between the produced poses $\widetilde{Y} = \{\tilde{y}_u\}_{u=1}^{U}$ and the ground truth $Y = \{y_u\}_{u=1}^{U}$:

$$\mathcal{L}_{MAE} = \frac{1}{U}\sum_{U}^{u=1} |\tilde{y}_u - y_u| \tag{11}$$

3.3 Multi-Hypothesis Aggregation Based Inference

Previous SLP methods [21, 43, 53] pay more attention to generating single hypotheses of sign poses, focusing little on aggregating multiple hypotheses to generate a single, high-confidence sign pose video. To explore the scalability of our GCDM framework in multi-hypothesis prediction, we introduce a multi-hypothesis aggregation mechanism in the reverse diffusion process. Specifically, we sample P sets of noise from a Gaussian distribution during the inference phase and feed them all into the pose denoiser. Driven by the gloss condition, multiple noisy pose streams are independently learned to obtain P hypotheses of the sign pose sequence. Finally, the GCDM outputs a single video of the sign poses by averaging multiple hypotheses. Combined with the Equations 10, this process can be expressed as:

$$\begin{aligned} &\widetilde{Y} = ReverseDiff(\widetilde{X}, \widetilde{H}_{1:P}) \Leftrightarrow \\ &\begin{cases} \widetilde{Y}_p = PoseDenoiser(\widetilde{X}, \widetilde{H}_p), p \in [1,P]; \\ \widetilde{Y} = Mean(\widetilde{Y}_{1:P}). \end{cases} \end{aligned} \tag{12}$$

It is noted that our method while utilizing the averaging approach for simplicity, recognizes the need for further refinement. Future work in this area could focus on developing more advanced aggregation strategies that combine the benefits of simplicity with the accuracy afforded by more complex techniques.

4 EXPERIMENTS

4.1 Experimental setup

4.1.1 Dataset. Following existing works [21, 43], we evaluate the proposed method on the dataset RWTH-PHOENIX-Weather2014T (PHOENIX14T) [3], a publicly available German sign language corpus, which provides 8257 parallel samples containing spoken sentences, sign glosses and sign videos. Specifically, the corpus covers 2887 different German words and 1066 different glosses, which is a challenging dataset due to the low video quality.

Table 2. Quantitative results on PHOENIX14T dataset. '†' indicates the reconstructed results.

Methods	DEV							TEST						
	B-1	B-2	B-3	B-4	ROUGE	WER↓	DTW-P↓	B-1	B-2	B-3	B-4	ROUGE	WER↓	DTW-P↓
Ground Truth	29.77	20.21	15.16	12.13	29.60	74.17	0.00	29.76	20.12	14.93	11.93	28.98	71.94	0.00
PT-base† [43]	9.53	3.45	1.62	0.72	8.61	98.53	29.33	9.47	3.37	1.47	0.59	8.88	98.36	28.48
PT-FP&GN† [43]	12.51	6.50	4.76	3.88	11.87	96.85	11.75	13.35	7.29	5.33	4.31	13.17	96.50	11.54
NAT-AT [21]	–	–	–	–	–	–	–	14.26	9.93	7.11	5.53	18.72	88.15	–
NAT-EA [21]	–	–	–	–	–	–	–	15.12	10.45	7.99	6.66	19.43	82.01	–
DET [58]	17.25	10.17	7.04	5.32	17.85	–	–	17.18	10.39	7.39	5.76	17.64	–	–
GEN [53]	18.86	11.10	7.68	5.77	19.43	90.34	11.94	18.71	11.53	8.09	6.20	19.79	90.37	11.89
SignVQNet [23]	–	–	–	6.77	–	–	–	–	–	–	6.88	–	–	–
GCDM (Ours)	**22.88**	**14.28**	**10.01**	**7.64**	**23.35**	82.81	**11.18**	**22.03**	**14.21**	**10.16**	**7.91**	**23.20**	**81.94**	**11.10**

4.1.2 Evaluation metrics. Following the widely used evaluation scheme in SLP [21, 22, 43, 46, 58], we use the classical SLT framework (*i.e.*, NSLT [3]) as a back-translation evaluation model, whose inputs are modified as the sign language pose sequences. To the best of our knowledge, there is currently no publicly available pre-trained back-translation evaluation model, so we retrained NSLT on PHOENIX14T referring to [21, 22]. We translate the generated poses back into sign gloss sequences and spoken sentences, and then calculate *BLEU*, *ROUGE* and Word Error Rate (*WER*) to measure the quality of produced sign poses. We provide *BLEU* n-grams from 1 to 4 for completeness.

Additionally, we provide the results of *DTW-P* for evaluating the quality of the generated sequences, which measures the sequential similarity between predicted pose sequence and ground truth based on Dynamic Time Warping (DTW).

4.1.3 Implementation details. We use OpenPose [4] to extract 2D joint coordinates from the original videos, and apply a skeleton model improvement estimation algorithm to convert the 2D coordinates into 3D sign poses, referring to [67]. In this work, we regard the transformed 3D pose sequences as ground-truth poses. All the transformer-based models in our GCDM are built with 2 layers, 4 heads and an embedding size of 512 (*i.e.*, $K = 2$, $M = 4$, $d_x = d_y = d_q = d_k = 512$). During the training of the GCDM, we apply Gaussian noise with a noise rate of 5, and set λ to 1.0 for simplicity. The parameter of the model is optimized with ADAM [27] optimizer and the learning rate is set to 1×10^{-3}. Experiments are performed with PyTorch on NVIDIA GeForce RTX 2080 Ti GPU.

4.2 Comparison with State-of-the-Arts

We compare our GCDM with state-of-the-art methods as follows:

- **PT-base** [43] proposes a pure transformer-based approach for SLP, which contains a symbolic transformer and a progressive transformer to translate spoken language into sign glosses and generate sign poses from glosses.
- **PT-FP&GN** [43] is an extension of PT(base) that introduces a future prediction mechanism (*i.e.*, predicting the next several frames from the current time step) and Gaussian noise for data augmentation.
- **NAT-AT** [21] first predicts the duration of the poses, and then utilizes a non-autoregressive model with a spatial-temporal graph convolutional pose generator to produce a sequence of sign language poses.
- **NAT-EA** [21] proposes a purely non-autoregressive model to directly predict sign poses, and explores the monotonic alignment between gloss feature sequences and pose sequences through an external aligner.

Table 3. Ablation studies of timesteps on PHOENIX14T dataset (Proposals $P = 1$).

Timesteps	DEV						TEST					
	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	ROUGE↑	WER↓	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	ROUGE↑	WER↓
$T = 10$	19.57	11.76	7.92	5.74	20.42	84.28	19.61	11.80	7.85	5.58	20.91	84.32
$T = 100$	20.80	12.91	9.05	6.89	22.23	83.56	20.05	12.62	8.86	6.81	21.46	82.55
$T = 1000$	**22.63**	**14.19**	**10.11**	**7.71**	**23.66**	**82.84**	**21.44**	**13.90**	**10.00**	**7.71**	**22.78**	**81.69**

- **DET** [58] designs a dual encoder transformer for SLP, which captures information from text and gloss to generate sign poses with facial landmarks and facial action units.
- **GEN** [53] aims to introduce the special token into gloss encoding to perform aggregate learning on the whole semantics of the gloss sequence, thereby enhancing the guidance ability of gloss semantics in the process of sign language generation.
- **SignVQNet** [23] presents the sign language vector quantization network, which leverages vector quantization to derive discrete representations from sign poses and integrates latent-level alignment for enhanced linguistic coherence in sign language production.

As shown in Table 2, SET-OBT performs prominent superior to all the other methods. These results reveal four points: (1) The GCDM method outperforms all other methods across all BLEU metrics on both the DEV and TEST sets. It achieves the highest scores with BLEU-1 at 22.88 on DEV and 22.03 on TEST, indicating its superior performance in matching the most frequent words. The GCDM also leads in ROUGE scores, with 23.35 on DEV and 23.20 on TEST, suggesting that it captures a greater extent of the reference sequences than the other models. For WER, GCDM shows the lowest (which indicates better performance) rates of 82.81 on DEV and 81.94 on TEST, surpassing the other methods. This points to GCDM's ability to accurately generate sign language poses that translate well back to spoken sentences. (2) The DTW-P metric, which measures the alignment of the generated pose sequence with the ground truth, is lowest for GCDM at 11.18 on DEV and 11.10 on TEST, demonstrating its precision in pose generation. (3) Among the other methods, GEN shows competitive results, especially in DEV, and PT-FP&GN performs notably well in the DTW-P metric on the DEV set, although it does not reach the performance level of GCDM. (4) Ground Truth scores provide a reference for the maximum achievable performance, showing that while GCDM is the leading method, there is still a gap between generated results and the Ground Truth.

In conclusion, the proposed GCDM establishes a new state-of-the-art performance on the PHOENIX14T dataset. These results demonstrate the effectiveness of GCDM in capturing the nuances of sign language, indicating its potential for practical applications in sign language translation and synthesis.

4.3 Ablation Study

4.3.1 Impact of Timesteps in Diffusion Model. Table 3 presents the results of ablation studies on the PHOENIX14T dataset. The studies investigate the impact of varying timesteps on the quality of the generated sign language videos. Observations from the table indicate that an increase in timesteps leads to improvements across all metrics in both the DEV and TEST sets. This suggests that more noise cycles enhance the model's ability to refine and generate higher-quality sign language pose videos.

The most significant improvements are noticeable when the timestep count is raised from 10 to 100. For instance, on the DEV set, BLEU-1 increases from 19.57 to 20.80, and WER decreases from 84.28 to 83.56, indicating higher precision and fewer errors, respectively. Diminishing returns are observed when further increasing timesteps from 100 to 1000. While there are still improvements, such as BLEU-1 rising from 20.80 to 22.63 in the DEV set

Table 4. Ablation studies of proposals on PHOENIX14T dataset (Timesteps $T = 100$).

Proposals	DEV						TEST					
	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	ROUGE↑	WER↓	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	ROUGE↑	WER↓
$P = 1$	20.80	12.91	9.05	6.89	22.23	83.56	20.05	12.62	8.86	6.81	21.46	82.55
$P = 2$	21.84	13.62	9.54	7.24	22.75	82.87	20.54	13.15	9.33	7.18	21.90	82.63
$P = 4$	21.48	13.67	9.93	**7.73**	22.73	84.04	21.04	13.51	9.59	7.39	22.16	82.93
$P = 6$	22.07	13.74	9.76	7.47	22.73	**82.44**	20.96	13.56	9.79	7.60	22.52	**81.83**
$P = 8$	**22.88**	**14.28**	**10.01**	7.64	**23.35**	82.81	**22.03**	**14.21**	**10.16**	**7.91**	**23.20**	81.94

and WER slightly dropping from 83.56 to 82.84, the increments are less pronounced compared to the previous increase. Comparing DEV and TEST sets, the patterns of improvement are consistent across both, although the TEST set generally exhibits slightly lower performance scores, indicating a robust model that generalizes well but with expected dips in an unseen environment.

The highest BLEU-1 score is achieved at T=1000 with 22.63 on the DEV set and 21.44 on the TEST set, suggesting that the model's ability to generate accurate first-word matches is better with more timesteps. The WER, an important metric for evaluating the coherent semantics of generating sign language videos, is lowest (indicating better performance) at T=1000 for both sets, at 82.84 for DEV and 81.69 for TEST.

In conclusion, the diffusion model's performance in translating generated sign language pose videos into glosses or text improves as the number of timesteps increases, with a notable leap from 10 to 100 timesteps and more gradual improvements after that. The consistency across metrics and datasets reinforces the model's reliability and potential utility for sign language translation tasks.

4.3.2 Number of Proposals in Multi-hypothesis Aggregation. Table 4 provides the results of an ablation study that assesses the performance impact of varying the number of proposals in the multi-hypothesis aggregation. Here, "proposals" refer to the number of hypotheses before aggregation, and the study employs a direct averaging method for multi-hypothesis aggregation. To focus on the effect of the number of proposals, the diffusion model's timesteps are consistently set to 100. From the results, we can deduce the following:

On the DEV set, there is a noticeable incremental improvement in BLEU-1 scores as the number of proposals increases, starting from 20.80 with P = 1 and peaking at 22.88 with P = 8. This indicates that the precision of matching the most frequent words improves with more proposals. The WER shows a decreasing trend (which indicates better performance) as the number of proposals increases. On the DEV set, the WER starts at 83.56 with P = 1 and decreases to 82.81 with P = 8, while on the TEST set, it goes from 82.55 to 81.94.

On the TEST set, similar to the DEV set, the performance metrics show improvements with an increasing number of proposals. For example, the BLEU-1 score rises from 20.05 with P = 1 to 22.03 with P = 8, and the WER drops from 82.55 to 81.94. The consistency of improvement in both DEV and TEST sets indicates that the benefits of using more proposals are robust and generalizable to unseen data.

It is worth noting that while the improvements are consistent, the rate of improvement appears to diminish as the number of proposals increases. This can be seen where the difference in BLEU-1 between P = 1 and P = 2 is more significant than between P = 6 and P = 8.

In conclusion, the results show that using a larger number of proposals in the multi-hypothesis aggregation approach enhances the performance of the diffusion model across all evaluated metrics. It suggests that integrating multiple hypotheses before averaging leads to a more accurate representation of sign language poses, which translates to better reverse translation quality from video to text or glosses. However, the diminishing returns

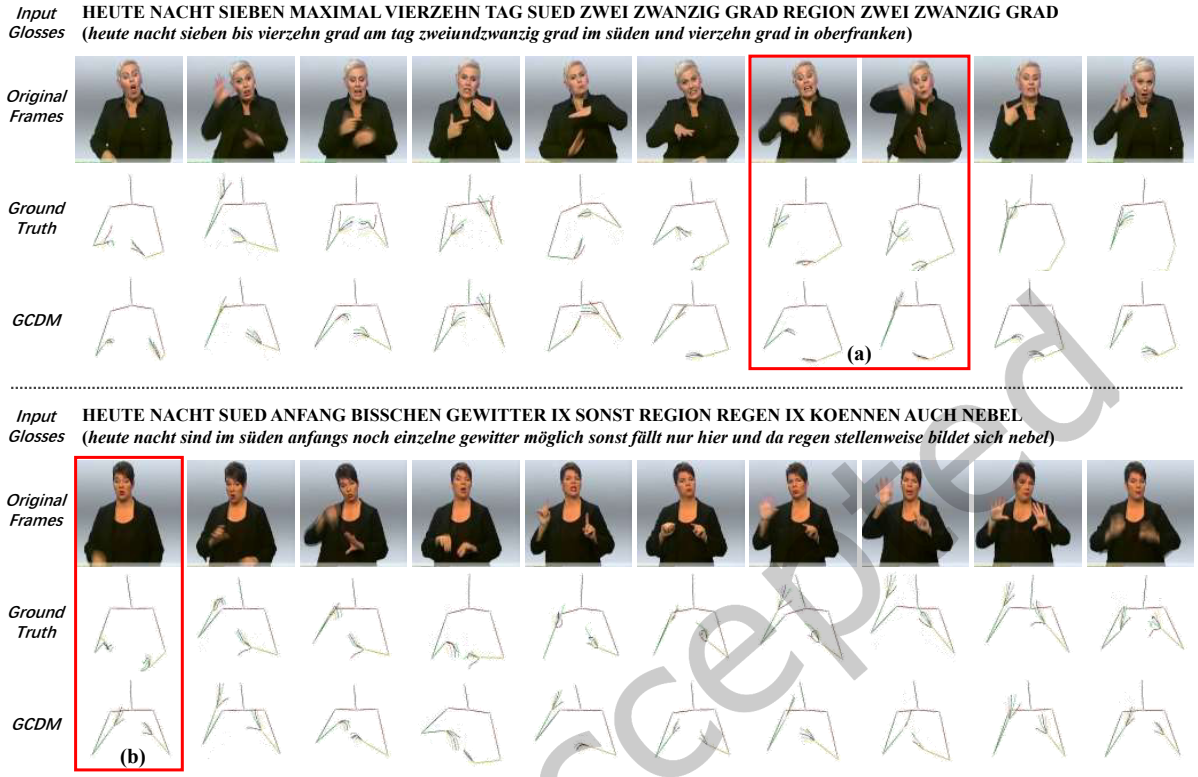

Fig. 3. Visual results of our GCDM on the benchmark. Challenging scenarios including fast motion and limbs not appearing on the frame are involved.

observed with higher proposal counts hint at a trade-off between computational resources and performance gains, a factor to consider in practical applications.

4.4 Qualitative Results

4.4.1 Visualization in Different Challenging Scenarios. Figure 3 offers a comprehensive visualization of the GCDM performance on the benchmark under different challenging scenarios, including rapid movements and instances where arms are not fully visible within the frame. In Figure 3 (a), the GCDM's ability to accurately capture and reproduce the gloss semantics in the generated poses is showcased. Despite instances where the ground truth data does not adequately capture finger details, potentially due to motion-blurring effects, the GCDM exhibits a remarkable capability to predict clear and distinct outcomes. This proficiency in maintaining definition under rapid motion conditions is a testament to the model's robustness. Figure 3 (b) further elucidates the GCDM's strength in generating lifelike poses by leveraging temporal dependencies, particularly when the subject's arm is not visible within the original frame. This demonstrates the model's advanced inferencing capabilities, where it effectively utilizes contextual information from previous and subsequent frames to reconstruct poses that are absent from the immediate frame under consideration.

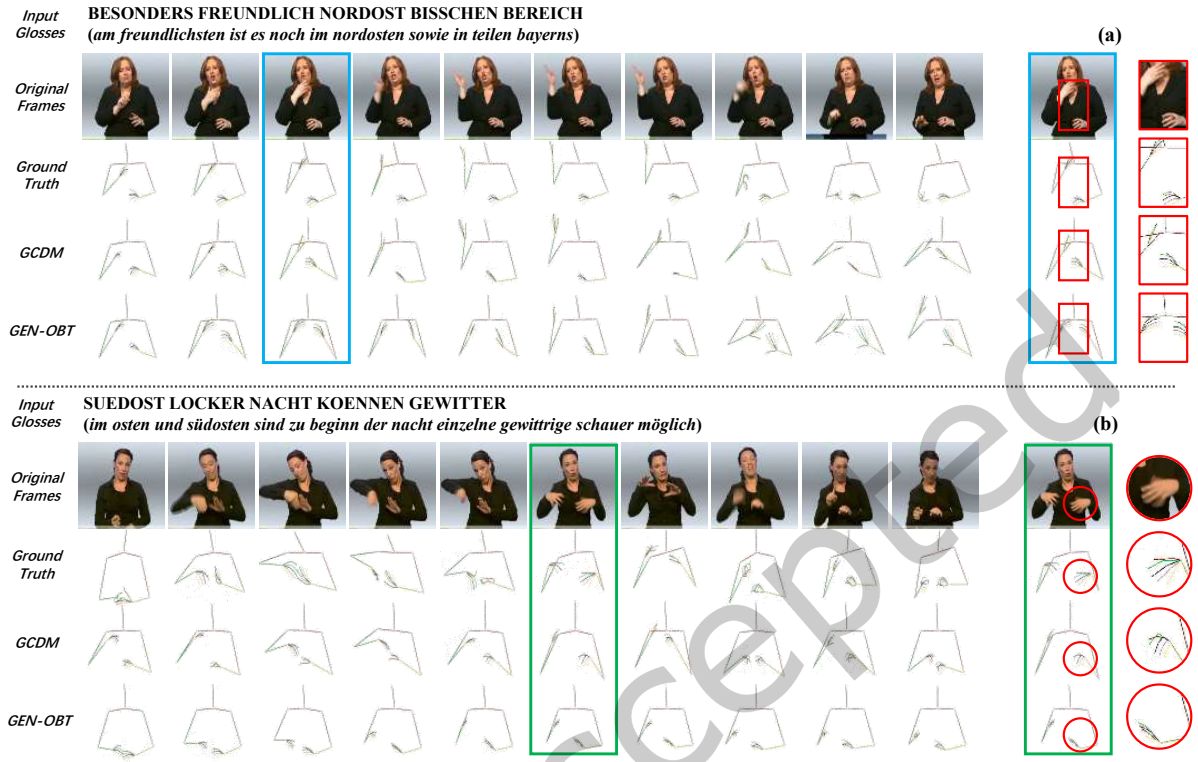

Fig. 4. Visualization of comparison between our GCDM and the existing method (*i.e.*, GEN-OBT [53]) on the benchmark.

4.4.2 Visualization Compared to Other Methods. We further provide two visual comparisons of the proposed GCDM against the existing GEN-OBT [53] in the SLP task. Ground truth annotations and the original video frames are attached to offer a benchmark for evaluation. In Figure 4 (a), the efficacy of GCDM in predicting hand positioning is accentuated. It exhibits a more precise reconstruction of the spatial relations between two hands, which is pivotal in sign language interpretation as it can significantly affect semantic conveyance. The red boxes underscore instances where GCDM markedly outperforms GEN-OBT, adhering more closely to the ground truth and thereby preserving the integrity of the sign representation. Figure 4 (b) scrutinizes the fidelity of finger detail rendering by both methods. The GCDM approach demonstrates superior definition in the articulation of finger postures, an aspect critical to the granularity of sign language. The red circles highlight the GCDM's enhanced clarity in finger positioning, whereas the GEN-OBT method's renderings appear comparatively indistinct, potentially leading to semantic discrepancies.

5 CONCLUSIONS

In this work, we propose a novel Gloss-driven Conditional Diffusion Model (GCDM) for Sign Language Production (SLP). The proposed GCDM is based on a diffusion model architecture, in which the sign gloss sequence is encoded by a Transformer-based encoder and integrated into the Gaussian noise in the pose denoiser as a semantic prior condition. Subsequently, the model converts the Gaussian noise with the gloss condition into sign language pose sequences through T-round denoising steps. Besides, a multi-hypothesis aggregation mechanism is introduced

in the inference phase, which generates multiple sign language pose sequence hypotheses and outputs a high-confidence sign video by averaging multiple hypotheses. Extensive experiments validate the effectiveness and robustness of the proposed method.

ACKNOWLEDGMENTS

This research is supported in part by grants from the National Natural Science Foundation of China (Grants No. 62272144, U20A20183, 61932009, 62302142, 62020106007), the Fundamental Research Funds for the Central Universities (Grants No. JZ2023HGQA0097), and China Postdoctoral Science Foundation (Grants No. 2022M720981).

REFERENCES

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. ViViT: A Video Vision Transformer. In *IEEE International Conference on Computer Vision*. 6836–6846.
[2] Qian Bao, Wu Liu, Jun Hong, Lingyu Duan, and Tao Mei. 2020. Pose-native Network Architecture Search for Multi-person Human Pose Estimation. In *ACM International Conference on Multimedia*. 592–600.
[3] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural Sign Language Translation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 7784–7793.
[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition*. 7291–7299.
[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end Object Detection with Transformers. In *European Conference on Computer Vision*. Springer, 213–229.
[6] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. 2021. WaveGrad: Estimating Gradients for Waveform Generation. In *International Conference on Learning Representations*.
[7] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research* 24, 240 (2023), 1–113.
[8] Runpeng Cui, Zhong Cao, Weishen Pan, Changshui Zhang, and Jianqiang Wang. 2019. Deep Gesture Video Generation with Learning on Regions of Interest. *IEEE Transactions on Multimedia* 22, 10 (2019), 2551–2563.
[9] Mathieu De Coster, Mieke Van Herreweghe, and Joni Dambre. 2021. Isolated Sign Recognition from RGB Video Using Pose Flow and Self-attention. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3441–3450.
[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.
[11] JRW Glauert, Ralph Elliott, SJ Cox, Judy Tryggvason, and Mary Sheard. 2006. VANESSA – A System for Communication Between Deaf and Hearing People. *Technology and Disability* 18, 4 (2006), 207–216.
[12] Dan Guo, Kun Li, Bin Hu, Yan Zhang, and Meng Wang. 2024. Benchmarking Micro-action Recognition: Dataset, Methods, and Applications. *IEEE Transactions on Circuits and Systems for Video Technology* (2024). https://doi.org/10.1109/TCSVT.2024.3358415
[13] Dan Guo, Shengeng Tang, Richang Hong, and Meng Wang. 2021. Review of Sign Language Recognition, Translation and Generation. *Computer Science* 48, 3 (2021), 60–70.
[14] Dan Guo, Shengeng Tang, Richang Hong, and Meng Wang. 2021. Sign Language Recognition. *Multimedia for Accessible Human Computer Interfaces* (2021), 23–59.
[15] Dan Guo, Shengeng Tang, and Meng Wang. 2019. Connectionist temporal modeling of video and language: a joint model for translation and sign labeling. In *International Joint Conference on Artificial Intelligence*. 751–757.
[16] Dan Guo, Shuo Wang, Qi Tian, and Meng Wang. 2019. Dense temporal convolution network for sign language translation. In *International Joint Conference on Artificial Intelligence*. 744–750.
[17] Dan Guo, Wengang Zhou, Anyang Li, Houqiang Li, and Meng Wang. 2019. Hierarchical recurrent deep fusion using adaptive clip summarization for sign language translation. *IEEE Transactions on Image Processing* 29 (2019), 1575–1590.
[18] Dan Guo, Wengang Zhou, Houqiang Li, and Meng Wang. 2017. Online Early-late Fusion Based on Adaptive HMM for Sign Language Recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications* 14, 1 (2017), 1–18.
[19] Dan Guo, Wengang Zhou, Houqiang Li, and Meng Wang. 2018. Hierarchical LSTM for sign language translation. In *AAAI Conference on Artificial Intelligence*, Vol. 32.
[20] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. 2018. Evaluation of speech-to-gesture generation using bi-directional LSTM network. In *International Conference on Intelligent Virtual Agents*. 79–86.

[21] Wencan Huang, Wenwen Pan, Zhou Zhao, and Qi Tian. 2021. Towards Fast and High-Quality Sign Language Production. In *ACM International Conference on Multimedia*. 3172–3181.
[22] Euijun Hwang, Jung-Ho Kim, and Jong-Cheol Park. 2021. Non-Autoregressive Sign Language Production with Gaussian Space. In *British Machine Vision Conference*.
[23] Eui Jun Hwang, Huije Lee, and Jong C Park. 2023. Autoregressive Sign Language Production: A Gloss-Free Approach with Discrete Representations. *arXiv preprint arXiv:2309.12179* (2023).
[24] Kostas Karpouzis, George Caridakis, S-E Fotinea, and Eleni Efthimiou. 2007. Educational Resources and Implementation of a Greek Sign Language Synthesis Architecture. *Computers & Education* 49, 1 (2007), 54–74.
[25] Dilek Kayahan and Tunga Güngör. 2019. A Hybrid Translation System from Turkish Spoken Language to Turkish Sign Language. In *International Symposium on INnovations in Intelligent SysTems and Applications*. 1–6.
[26] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *IEEE Conference on Computer Vision and Pattern Recognition*. 15954–15964.
[27] Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. (2015).
[28] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2021. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *International Conference on Learning Representations*.
[29] Dimitris Kouremenos, Klimis S Ntalianis, Giorgos Siolas, and Andreas Stafylopatis. 2018. Statistical Machine Translation for Greek to Greek Sign Language Using Parallel Corpora Produced via Rule-Based Machine Translation. In *International Conference on Tools with Artificial Intelligence*. 28–42.
[30] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. 2019. Pifpaf: Composite Fields for Human Pose Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 11977–11986.
[31] Shyam Krishna and Janmesh Ukey. 2021. GAN Based Indian Sign Language Synthesis. In *Indian Conference on Vision, Graphics and Image Processing*. 1–8.
[32] Chen Li and Gim Hee Lee. 2019. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *IEEE Conference on Computer Vision and Pattern Recognition*. 9887–9895.
[33] Kun Li, Dan Guo, and Meng Wang. 2021. Proposal-free video grounding with contextual pyramid network. In *AAAI Conference on Artificial Intelligence*, Vol. 35. 1902–1910.
[34] Kun Li, Dan Guo, and Meng Wang. 2023. ViGT: proposal-free video grounding with a learnable token in the transformer. *Science China Information Sciences* 66, 10 (2023), 202102.
[35] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. 2022. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 13147–13156.
[36] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. 2023. VideoFusion: Decomposed Diffusion Models for High-Quality Video Generation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 10209–10218.
[37] Taro Miyazaki, Yusuke Morita, and Masanori Sano. 2020. Machine Translation from Spoken Language to Sign Language Using Pre-trained Language Model As Encoder. In *Workshop on the Representation and Processing of Sign Languages*. 139–144.
[38] B Natarajan and R Elakkiya. 2022. Dynamic GAN for High-Quality Sign Language Video Generation from Skeletal Poses Using Generative Adversarial Networks. *Soft Computing* (2022).
[39] Gabriele Prato, Ella Charlaix, and Mehdi Rezagholizadeh. 2020. Fully Quantized Transformer for Machine Translation. In *Conference on Empirical Methods in Natural Language Processing*. 1–14.
[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.
[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition*. 10684–10695.
[42] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. Adversarial Training for Multi-Channel Sign Language Production. In *British Machine Vision Conference*.
[43] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. Progressive Transformers for End-to-end Sign Language Production. In *European Conference on Computer Vision*. 687–705.
[44] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021. AnonySign: Novel Human Appearance Synthesis for Sign Language Video Anonymisation. In *IEEE International Conference on Automatic Face and Gesture Recognition*. 1–8.
[45] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021. Continuous 3d Multi-channel Sign Language Production Via Progressive Transformers and Mixture Density Networks. *International Journal of Computer Vision* 129, 7 (2021), 2113–2135.
[46] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021. Mixed SIGNals: Sign Language Production via a Mixture of Motion Primitives. In *IEEE International Conference on Computer Vision*. 1919–1929.

[47] Peipei Song, Dan Guo, Xun Yang, Shengeng Tang, and Meng Wang. 2024. Emotional Video Captioning With Vision-Based Emotion Interpretation Network. *IEEE Transactions on Image Processing* 33 (2024), 1122–1135.
[48] Peipei Song, Dan Guo, Xun Yang, Shengeng Tang, Erkun Yang, and Meng Wang. 2023. Emotion-Prior Awareness Network for Emotional Video Captioning. In *ACM International Conference on Multimedia*. 589–600.
[49] Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. 2020. Text2Sign: Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks. *International Journal of Computer Vision* 128, 4 (2020), 891–908.
[50] Stephanie Stoll, Simon Hadfield, and Richard Bowden. 2020. SignSynth: Data-Driven Sign Language Video Generation. In *European Conference on Computer Vision*. 353–370.
[51] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *International Conference on Learning Representations*.
[52] Shengeng Tang, Dan Guo, Richang Hong, and Meng Wang. 2022. Graph-based multimodal sequential embedding for sign language translation. *IEEE Transactions on Multimedia* 24 (2022), 4433–4445.
[53] Shengeng Tang, Richang Hong, Dan Guo, and Meng Wang. 2022. Gloss semantic-enhanced network with online back-translation for sign language production. In *ACM International Conference on Multimedia*. 5630–5638.
[54] Inigo Jauregi Unanue, Jacob Parnell, and Massimo Piccardi. 2021. BERTTune: Fine-Tuning Neural Machine Translation with BERTScore. In *Annual Meeting of the Association for Computational Linguistics*. 915–924.
[55] Neel Vasani, Pratik Autee, Samip Kalyani, and Ruhina Karani. 2020. Generation of Indian Sign Language by Sentence Processing and Generative Adversarial Networks. In *International Conference on Information Systems Security*. 1250–1255.
[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. *Advances in Neural Information Processing Systems* 30 (2017).
[57] Lucas Ventura, Amanda Duarte, and Xavier Giró-i Nieto. 2020. Can Everybody Sign Now? Exploring Sign Language Video Generation from 2d Poses. In *Sign Language Recognition, Translation & Production*.
[58] Carla Viegas, Mert Inan, Lorna Quandt, and Malihe Alikhani. 2023. Including Facial Expressions in Contextual Embeddings for Sign Language Generation. In *Joint Conference on Lexical and Computational Semantics*. 1–10.
[59] Fei Wang, Dan Guo, Kun Li, and Meng Wang. 2024. Eulermormer: Robust eulerian motion magnification via dynamic filtering within transformer. In *AAAI Conference on Artificial Intelligence*, Vol. 38. 5345–5353.
[60] Fei Wang, Dan Guo, Kun Li, Zhun Zhong, and Meng Wang. 2024. Frequency Decoupling for Motion Magnification via Multi-Level Isomorphic Architecture. *arXiv preprint arXiv:2403.07347* (2024).
[61] Shuo Wang, Dan Guo, Wen-gang Zhou, Zheng-Jun Zha, and Meng Wang. 2018. Connectionist temporal fusion for sign language translation. In *ACM International Conference on Multimedia*. 1483–1491.
[62] Xinshuo Weng, Boris Ivanovic, and Marco Pavone. 2022. Mtp: Multi-hypothesis tracking and prediction for reduced error propagation. In *IEEE Intelligent Vehicles Symposium*. IEEE, 1218–1225.
[63] Qinkun Xiao, Minying Qin, and Yuting Yin. 2020. Skeleton-based Chinese Sign Language Recognition and Generation for Bidirectional Communication Between Deaf and Hearing People. *Neural Networks* 125 (2020), 41–55.
[64] Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Weinan Zhang, Yong Yu, and Lei Li. 2020. Towards Making the Most of Bert in Neural Machine Translation. In *AAAI Conference on Artificial Intelligence*, Vol. 34. 9378–9385.
[65] Ling Yang, Jingwei Liu, Shenda Hong, Zhilong Zhang, Zhilin Huang, Zheming Cai, Wentao Zhang, and Bin CUI. 2024. Improving Diffusion-Based Image Synthesis with Context Prediction. In *Neural Information Processing Systems*, Vol. 36. 37636–37656.
[66] Aoxiong Yin, Zhou Zhao, Jinglin Liu, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. 2021. SimulSLT: End-to-End Simultaneous Sign Language Translation. In *ACM International Conference on Multimedia*. 4118–4127.
[67] Jan Zelinka and Jakub Kanis. 2020. Neural Sign Language Synthesis: Words Are Our Glosses. In *International Workshop on Applications of Computer Vision*. 3395–3403.
[68] Jan Zelinka, Jakub Kanis, and Petr Salajka. 2019. NN-based Czech Sign Language Synthesis. In *International Conference on Speech and Computer*. 559–568.
[69] Jiali Zeng, Shuangzhi Wu, Yongjing Yin, Yufan Jiang, and Mu Li. 2021. Recurrent Attention for Neural Machine Translation. In *Conference on Empirical Methods in Natural Language Processing*. 3216–3225.
[70] Ni Zeng, Yiqiang Chen, Yang Gu, Dongdong Liu, and Yunbing Xing. 2020. Highly Fluent Sign Language Synthesis Based on Variable Motion Frame Interpolation. In *IEEE International Conference on Systems, Man, and Cybernetics*. 1772–1777.
[71] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *IEEE International Conference on Computer Vision*. 3836–3847.
[72] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068* (2022).
[73] Tianfu Zhang, He-Yan Huang, Chong Feng, and Longbing Cao. 2021. Enlivening Redundant Heads in Multi-head Self-attention for Machine Translation. In *Conference on Empirical Methods in Natural Language Processing*. 3238–3248.

[74] Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. 2021. Spatial-temporal Multi-cue Network for Sign Language Recognition and Translation. *IEEE Transactions on Multimedia* (2021).
[75] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2020. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *International Conference on Learning Representations*.

Received 23 February 2024; revised 3 April 2024; accepted 24 April 2024