

# Multi-Channel Spatio-Temporal Transformer for Sign Language Production

Xiaohan Ma<sup>1</sup>, Rize Jin<sup>2</sup>, Tae-Sun Chung<sup>1</sup>

<sup>1</sup> Department of Artificial Intelligence, Ajou University

<sup>2</sup> School of Software, Tiangong University

{maxiaohan, tschung}@ajou.ac.kr, jinrize@tiangong.edu.cn

## Abstract

The task of Sign Language Production (SLP) in machine learning involves converting text-based spoken language into corresponding sign language expressions. Sign language conveys meaning through the continuous movement of multiple articulators, including manual and non-manual channels. However, most current Transformer-based SLP models convert these multi-channel sign poses into a unified feature representation, ignoring the inherent structural correlations between channels. This paper introduces a novel approach called MCST-Transformer for skeletal sign language production. It employs multi-channel spatial attention to capture correlations across various channels within each frame, and temporal attention to learn sequential dependencies for each channel over time. Additionally, the paper explores and experiments with multiple fusion techniques to combine the spatial and temporal representations into naturalistic sign sequences. To validate the effectiveness of the proposed MCST-Transformer model and its constituent components, extensive experiments were conducted on two benchmark sign language datasets from diverse cultures. The results demonstrate that this new approach outperforms state-of-the-art models on both datasets.

**Keywords:** Sign Language Production; Transformer; Spatio-Temporal Fusion

## 1. Introduction

The World Health Organization (WHO) estimates that approximately 5% of the global population is afflicted with hearing loss of moderate or greater intensity (World Health Organization, 2021). Although sign language is not ubiquitously adopted by this demographic, it serves as the predominant communication medium for deaf communities. In contrast to spoken language, sign language achieves communication through continuous movements across multiple channels, namely, the facial, upper body, and hands (Sutton-Spence and Woll, 1999; Bragg et al., 2019). Each channel plays a pivotal role in the expression of sign language (Bragg et al., 2019). Furthermore, the spatial configuration, position, and temporal motion of these channels collectively form the grammar and semantic structure of sign language (Stokoe, 2005).

In the domain of Sign Language Production (SLP), models are designed to generate sign language expressions from textual sequences, which are commonly represented as sequences of glosses or words. In SLP models, sign language can be represented in various forms, such as sign pose sequences (skeleton joint coordinates) (Zelinka and Kanis, 2020; Saunders et al., 2020a, 2021a, 2020b, 2021b; Hwang et al., 2021; Saunders et al., 2022b; Huang et al., 2021; Ma et al., 2024), animations (Zwitserslood et al., 2004; Glauert et al., 2006; Elliott et al., 2008; Ebling and Huenerfauth, 2015; McDonald et al., 2016; Cui et al., 2022), and photo-realistic videos (Stoll et al., 2018, 2020; Saunders et al., 2022a). This study

focuses on the generation of sign pose sequences from sequences of words or glosses, which is more suitable for downstream applications.

Owing to the differences in tokenization and the phonological characteristics between spoken and sign languages, SLP models encounter difficulties in accurately mapping simple textual inputs to continuous pose sequences that exhibit variations across multiple visual channels. A sequence of sign poses, encompassing continuous frames across several channels, not only the spatial relationships among different channels but also the inherent temporal dynamics specific to each channel. Recently developed Transformer-based SLP approaches (Zelinka and Kanis, 2020; Saunders et al., 2020b,a, 2021a,b) map individual frames of full-channel sign poses into abstract representations, leveraging attention mechanisms to discern temporal dependencies between full-channel frames, and subsequently generate the entire pose sequence from text. These approaches primarily focus on the temporal context and may miss spatial relationships across multiple channels within each frame's multi-channel sign poses. Concurrently, another avenue of research, namely, graph-based SLP models (Huang et al., 2021; Saunders et al., 2022b) adopt a "joint-level" approach, where joints serve as nodes, and limbs form the edges. By converting the sign pose into a graph structure, these networks leverage spatio-temporal convolutional or attention mechanisms to learn intra- and inter-frame correlations.

In this study, we introduce the Multi-Channel Spatio-Temporal Transformer (MCST-Transformer)

for sign language production, which is designed to operate at the channel level, aiming to discern the spatial collaborative interactions and capture the temporal dynamics among individual channels. Specifically, our model incorporates a spatial attention module dedicated to decoding multi-channel spatial inter-dependencies and temporal attention module to discern the temporal dynamics unique to each channel. Furthermore, to improve expressiveness and naturalness in the generated sign sequences, we evaluate three distinct fusion techniques aiming to effectively integrate insights from both spatial and temporal attention features. To confirm the adaptability and potency of MCST-Transformer, we conduct an extensive evaluation of the model using German and Korean sign language production benchmarks. The results highlight the advantages our model offers over the existing transformer-based approaches.

The key contributions of this study can be summarized as follows: **1)** We introduce a Multi-Channel Spatio-Temporal Transformer (MCST-Transformer) model, which is designed to exploit the spatial structural relationships between channels, as well as the temporal dependencies inherent in each channel in sign pose sequences. **2)** We explore three distinct fusion methods to effectively combine spatial and temporal attention features, ensuring a cohesive integration that leads to generating highly expressive and naturalistic full-channel sign language sequences. **3)** We conduct a comprehensive evaluation on two distinct sign language datasets: German and Korean sign language datasets, highlighting the robustness and effectiveness of the proposed model.

## 2. Related work

Previous conventional SLP models have used artificial avatars to display sign language using parameterized glosses (Zwitserslood et al., 2004; Glauert et al., 2006; Ebling and Huenerfauth, 2015; McDonald et al., 2016; Bangham et al., 2000; Efthimiou et al., 2012). These studies suffer from unnatural movements and missing non-manual information. In some studies, non-manual information, such as facial expressions, has been included (Elliott et al., 2008; Ebling and Glauert, 2013; Cox et al., 2002) but mouth patterns are difficult for human operators to handcraft (Kipp et al., 2011). Motion capture data can be used to enhance the natural appearance of animation (Gibbet et al., 2016); however, scalability is limited because of the costs associated with capturing and annotating the data.

Recently, deep learning models have been applied to SLP. Stoll et al. (2018, 2020) proposed the first approach for generating continuous sign language videos without relying on traditional graphi-

cal avatars, by integrating neural machine translation (NMT) techniques with a generative adversarial network (GAN). It first translates text sequences into gloss sequences utilizing NMT methods and then converts the resulting gloss information into 2D skeletal pose sequences via a trained lookup table. Subsequently, these pose sequences serve to condition a GAN model, producing photo-realistic sign language videos. Zelinka and Kanis (2020) first proposed a fully trainable end-to-end SLP model for generating fixed-length sign pose sequences. To extract high-quality sign pose sequences, they designed a gradient descent-based method to estimate 3D skeletons from 2D skeletons extracted using OpenPose (Cao et al., 2019). Saunders et al. (2020b) proposed the Pro-Transformer that generates manual and body skeletons using powerful feed-forward and recurrent transformers. To achieve more expressive and articulate production, Saunders et al. (2020a) combined the conditional GAN model with the Pro-Transformer (Saunders et al., 2020b) to extend sign production to non-manual features. More recently, a series of research proposed Pro-Transformer-based SLP models. For example, Saunders et al. (2021a) added constraints with mixture distribution, Saunders et al. (2021b) introduced a mixture of motion primitives to produce better full-channel sign pose sequences. Ma et al. (2024) introduced a cascade dual decoder Transformer model to enhance the hand details in the full-channel sign language production. These transformer-based SLP models tend to represent full-channel sign poses in a more abstract form, often compromising the intricate structural inter-connections among the channels. Concurrently, another avenue of research has ventured into representing sign pose sequences as graphs (Huang et al., 2021; Saunders et al., 2022b; Cui et al., 2022), where articulation points are considered as nodes and limb connections are considered as edges. These graph-based methods incorporate spatio-temporal convolutional or attention mechanisms to capture both intra- and inter-frame information correlations. However, these models disregard the distinct contributions and synergies of the individual channels in gesture expressions.

To learn the channel-level spatial and temporal relationships, we first introduce a multi-channel spatial attention mechanism tailored to capture the intricate spatial relationships between various sign channels and a dedicated temporal attention layer that discovers the temporal dependencies inherent to each channel. Moreover, we explore three distinct fusion methods to effectively merge spatial and temporal attention features, ensuring the cohesive interplay vital for the generation of sign sequences. Furthermore, in contrast to the previously mentioned studies focusing solely on a single sign

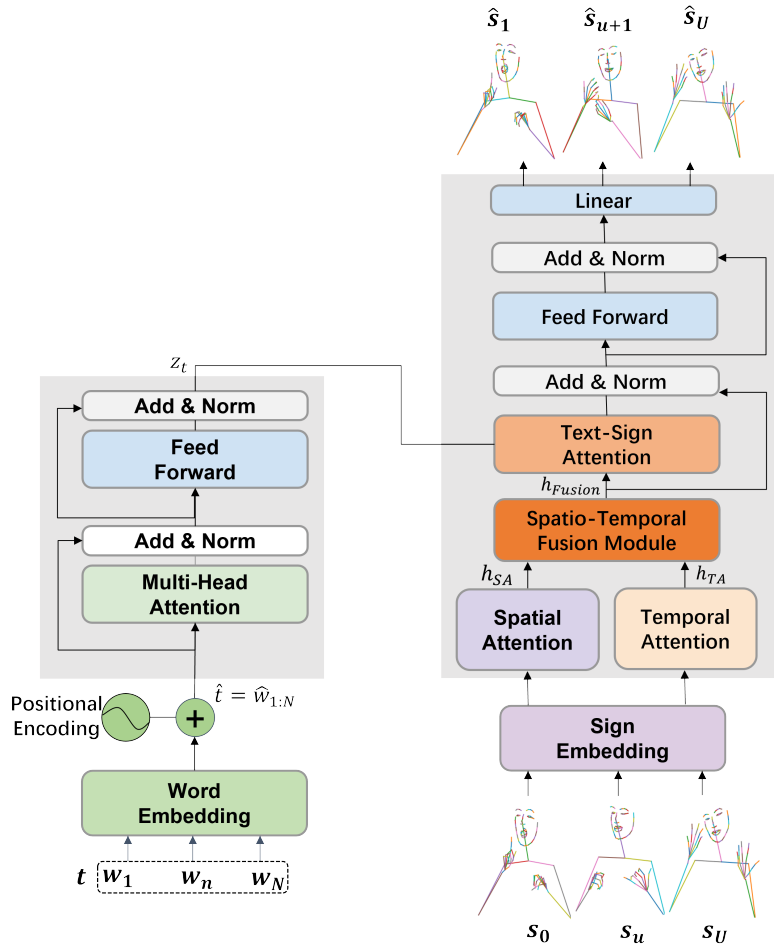


Figure 1: Overview of the proposed MCST-Transformer model, which is composed of an encoder and multi-channel spatio-temporal decoder. The former learns semantic representations from source sentences  $t$ , while the latter captures both spatial and temporal dependencies of sign pose sequences to generate full-articulatory sign text sequence  $\hat{s}_{1:U}$ . By convention, we add positional encoding (PE) into the word and spatial embeddings, respectively, preserving the orders of sequences.

language (Saunders et al., 2020a, 2021a, 2020b, 2021b; Hwang et al., 2021; Stoll et al., 2018, 2020), our model has the capability to generate sign pose sequences for both German and Korean, establishing a foundation for the production of multiple sign languages.

### 3. Methodology

In this section, we delve into the details of the proposed MCST-Transformer model, which is designed to capture the inherent spatial structures and temporal dynamics of sign pose sequences. We begin by clarifying the problem definition of SLP and identifying the challenges inherent to the foundational model, Pro-Transformer. Furthermore, to generate expressive and naturalistic full-channel sign language sequences, we exploit various distinct fusion methods to integrate both spatial and temporal attention features.

#### 3.1. Problem Definition

Given a source sequence  $t = \{w_1, \dots, w_n, \dots, w_N\}$  with  $N$  words and a full-channel sign sequence  $s = \{s_1, \dots, s_u, \dots, s_U\}$  with  $U$  frames, where each frame  $s_u$  represents a single sign language pose. Each sign language pose contains multi-channel features, which can be divided into three parts: the face (72 keypoints), the left upper body (including shoulder, arm, and fingers, totaling 24 keypoints), and the right upper body (24 keypoints). The goal of the SLP model is to learn the conditional probability  $P(s|t)$ .

#### 3.2. Pro-Transformer

The Pro-Transformer (Saunders et al., 2020b) utilizes the transformer model (Vaswani et al., 2017) to translate a gloss or word sequence into a sign pose sequence, which is essentially a sequence of 3D joint coordinates of the signer’s skeleton. Specifically, it consists of an encoder and decoder model,

where the former encodes the source sequence  $w_{1:N}$  to a contextual representation  $z_t$ , which can be formulated as follows:

$$z_t = \text{Encoder}_{\text{PT}}(\hat{w}_n | \hat{w}_{1:N}) \quad (1)$$

where  $\hat{w}_{1:N}$  represents the embedded representation of the source sequence  $w_{1:N}$ . The decoder follows an autoregressive approach to generate a sign pose frame  $\hat{s}_u$  at each time-step, along with the corresponding counter value  $c_u$ :

$$[\hat{s}_u, c_u] = \text{Decoder}_{\text{PT}}(\hat{j}_{1:u-1}, z_t) \quad (2)$$

where  $\hat{s}_u$  is the  $u$ -th frame of the produced sign pose sequence, given the text representation  $z_t$  and counter-concatenated joint embedding  $\hat{j}_{1:u-1}$  of the previous sign pose sequences  $s_{1:u-1}$ .  $c_u$  is the counter value of the  $u$ -th produced frame, which is obtained by the Counter Encoding in Pro-Transformer (Saunders et al., 2020b) to provide temporal information:

$$c_u = \frac{u}{U}, \quad c_u \in [0, 1] \quad (3)$$

The Pro-Transformer model roughly vectorizes full-channel sign pose and combines these abstract vectorizations with text representations to predict subsequent sign pose frames using attention operations. However, note that each full-channel sign pose contains multiple channel poses. The attention mechanism in Pro-Transformer operates on full-channel pose frames and primarily captures the temporal dependencies between frames, ignoring the structural dependencies of channels in each full-channel pose.

### 3.3. Multi-Channel Spatio-Temporal Transformer

To capture the spatial and temporal dependencies among channels for sign language production, this paper introduces a novel approach, MCST-Transformer, which incorporates temporal attention to capture dependencies of individual channels in the temporal dimension and spatial attention to capture structural interactions between different channels. In addition, we explore various fusion methods to combine the spatial and temporal representations. It should be noted that our model shows versatility and can be easily integrated into any backbone network, effectively assembling different representation methods of channel-level poses in a simple way. Figure 1 illustrates the overall architecture. Each component of the proposed model will be discussed in subsequent subsections.

#### 3.3.1. Encoder

The MCST-Transformer’s encoder utilizes stacked transformer encoder blocks with multi-head self-attention to acquire contextualized representations

capturing long-range dependencies within source sequences. Similar to the Pro-Transformer (Saunders et al., 2020b) model, the source sequence  $t = w_{1:N}$  is first embedded via a linear embedding layer and then added to the corresponding position embedding obtained using a predefined sinusoidal function. The embedded representation  $\hat{t} = \hat{w}_{1:N}$  of the source sequence is then fed into a stack of encoder blocks to construct the contextual representation  $z_t$ , which can be formulated as follows:

$$z_t = \text{Encoder}(\hat{w}_n | \hat{w}_{1:N}) \quad (4)$$

#### 3.3.2. Multi-Channel Spatio-Temporal Decoder

The aim of the multi-channel spatio-temporal decoder is to generate complete full-channel sign pose sequences. It operates by taking the text representation as input and auto-regressively predicting the full-channel sign sequence by leveraging spatial and temporal information learned from preceding sign poses. It comprises four modules: the temporal attention module, the spatial attention module, the spatio-temporal fusion module, and the text-sign attention module.

**Channel-Specific and Full-Channel Embeddings.** The individual full-channel sign pose, denoted as  $s_u$ , comprises a total of 120 keypoints, which are further categorized into three channels: the face  $s_u^f$  (encompassing the face and neck regions, totaling 72 keypoints), the left body  $s_u^l$ , and the right body  $s_u^r$  (comprising the shoulder, arm, and hand, with 24 keypoints on each side). To facilitate representation, we employ separate linear embedding layers and positional encoding techniques for these three channels and the full-channel sign pose. This entire embedding process can be formally expressed as follows:

$$E_u^i = PE_i = \text{PositionEncoding}(W^{(i,E)} \cdot s_u^i) \quad (5)$$

$$E_u^s = PE_s = \text{PositionEncoding}(W^{(s,E)} \cdot s_u) \quad (6)$$

where  $s_u^i$  can represent either  $s_u^f$ ,  $s_u^l$ , or  $s_u^r$ , whereas  $s_u$  refers to the full-channel pose. We utilize a predefined sinusoidal function to encode positional information for each of these poses.

**Multi-Channel Spatial Attention Module.** In Pro-Transformer, the attention modules operate on the entire full-channel embeddings  $E_u^s$ , implicitly capturing the relationships between the channels. To explicitly model the inter-channel dependencies of each single sign pose, we introduce a multi-channel spatial attention module, as illustrated in Figure 2. We apply two distinct linear transformations to the full-channel pose embedding  $E_u^s$  to derive the key and value components for the attention layer. Simultaneously, we compute the query by applying separate linear transformations to three distinct spatial

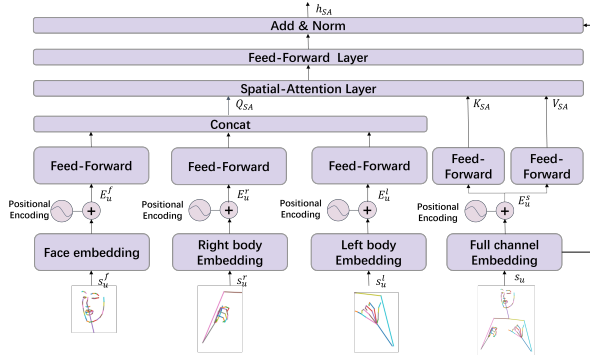


Figure 2: The architecture of multi-channel spatial attention module.

embeddings  $E_u^f$ ,  $E_u^l$ , and  $E_u^r$ , corresponding to the face, left body, and right body parts, respectively. We then concatenate the transformed queries from each channel to form the final query representation. Subsequently, the key, query, and value are passed through a multi-head self-attention layer to calculate the spatial relationships between each channel and the full-channel pose. The entire process can be formulated as follows:

$$Q_{SA} = \text{Concat}(W^{(f,Q)} \cdot E_u^f, W^{(l,Q)} \cdot E_u^l, W^{(r,Q)} \cdot E_u^r) \quad (7)$$

$$K_{SA} = W^{(s,K)} \cdot E_u^s \quad (8)$$

$$V_{SA} = W^{(s,V)} \cdot E_u^s \quad (9)$$

$$h_{SA} = \text{softmax}\left(\frac{Q_{SA} \cdot K_{SA}^T}{\sqrt{d_k}}\right)V_{SA} \quad (10)$$

where  $d_k$  is the dimensionality of the full-channel sign embedding. Note that we adjust the masking method in the spatial attention to ensure it concentrates only on relevant past inter-channel information.

**Multi-Channel Temporal Attention Module.** The multi-channel temporal attention module is designed to capture the temporal dynamics of each channel, which includes the face  $s^f$ , left body  $s^l$ , and right body  $s^r$ . The structure of the temporal attention module is visualized in Figure 3. First, we perform linear transformations on individual channel embeddings  $E^i$  to derive the respective keys  $K^i$ , queries  $Q^i$ , and values  $V^i$  as follows:

$$K^i = W^{(i,K)} \cdot E^i \quad (11)$$

$$Q^i = W^{(i,Q)} \cdot E^i \quad (12)$$

$$V^i = W^{(i,V)} \cdot E^i \quad (13)$$

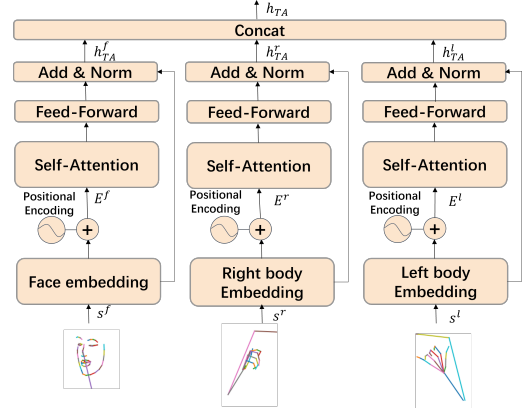


Figure 3: The architecture of multi-channel temporal attention module.

where  $W^{(i,K)}$ ,  $W^{(i,Q)}$ , and  $W^{(i,V)}$  are the learnable weight matrices for channel  $i$  and  $E^i$  refers to the embedding of the three channels  $E^f$ ,  $E^l$ , and  $E^r$ .

Next, the self-attention mechanism is applied to each channel, capturing their unique temporal dependencies. In order to prevent information leakage from subsequent frames, a dedicated mask is applied to each self-attention operation. Finally, the resulting individual temporal attention features are connected, thus producing fine-grained temporal patterns.

$$h_{TA}^i = \text{softmax}\left(\frac{Q^i \cdot (K^i)^T}{\sqrt{d_k^i}}\right)V^i \quad (14)$$

$$h_{TA} = \text{Concat}(h_{TA}^i) \quad (15)$$

where  $d_k^i$  is the feature dimensionality of channel  $i$ . **Spatio-Temporal Fusion Module.** In order to better integrate spatial and temporal features, we explore three fusion methods: 1) Parallel addition: Two attention modules operate concurrently, and then their resulting features are merged via addition. This approach ensures that both spatial and temporal features are given equal importance. 2) Sequential operation: The temporal attention module precedes the spatial module, with the output of the former serving as the input to the latter. This arrangement encourages that spatial decisions are guided by preceding temporal information, making it particularly suited for scenarios where the sequence of feature processing holds significant importance. 3) Gating fusion: We adopt a gating mechanism (Cui et al., 2022) to dynamically integrate the spatial and temporal features. This mechanism can be formally represented by:

$$\text{gate} = \sigma(W_1 \cdot h_{TA} + W_2 \cdot h_{SA} + b) \quad (16)$$

where  $W_1$  and  $W_2$  denote weight matrices, and  $b$  is the bias vector of the linear layer. The gate is subsequently used to weigh the spatial and temporal

attention features:

$$h_{\text{fusion}} = (1 - \text{gate}) \odot h_{SA} + \text{gate} \odot h_{TA} \quad (17)$$

Finally, the fusion representation is fed into the text-sign attention layer to align the resource and sign pose sequences. After the stacked decoder blocks, we project the output using a linear layer to generate the predicted sign pose sequence.

Our decoder is trained using the mean squared error (MSE) loss calculated between the predicted and ground-truth sign pose sequences.

$$\mathcal{L} = \frac{1}{U} \sum_{i=1}^U (s_i - \hat{s}_i)^2 \quad (18)$$

where  $\hat{s}_i$  and  $s_i$  refer to the frames of the produced and ground-truth sign pose sequences, respectively.

## 4. Experimental Setup

### 4.1. Datasets

To verify the robustness and scalability of the proposed approach, we conduct evaluations of our model using two sign language datasets, which include sequences of sign language and their corresponding translations into spoken language or sequences of glosses.

The first dataset, RWTH-PHOENIX-Weather 2014T (PHOENIX14T) (Camgoz et al., 2018), comprises German sign language interpretations of weather forecasts broadcasted by the German television station PHOENIX. This dataset encompasses 8,257 video segments performed by 9 signers, distributed into 7,096 for training, 519 for validation, and 642 for testing. It has a spoken language vocabulary of 2,887 words and a sign language vocabulary of 1,066. Annotations for each sign video include both glosses and German translations.

Furthermore, we evaluate our model on a large-scale Korean sign language dataset, the Korean Sign Language Guide Dataset (KSL-Guide) (Ham et al., 2021), which offers a rich collection of sign language sentences, individual words, and finger-spelled words and numbers tailored to transportation and navigation dialogues. The KSL-Guide encompasses 2,000 unique sentences, with each of the 20 signers recording videos for each sentence using a multi-view camera system with five cameras, and only the front view is used in this study. To ensure the diversity of the sign language skeleton, we choose three signer’s sign language video segments, totaling 6,000 sentence pairs. These are allocated into 5,400 for the training set, 300 for validation, and 300 for the test set.

The sign language data used for training our model are the sequences of 3D skeleton poses,

which are concatenated vectors of the 3D joint coordinates of the upper body skeleton. The PHOENIX14T dataset contains only sign video frames. Therefore, we utilize OpenPose (Cao et al., 2019) to extract 2D joint coordinates from the frames and then convert them to 3D joint coordinates ( $x$ ,  $y$ , and  $z$  coordinates of 120 joints) using a 3D-model-based skeletal model correction method (Zelinka and Kanis, 2020). The KSL-Guide dataset provides 3D joint coordinates for each sentence, which are utilized to extract 2D coordinates from two viewpoints and correct them by human annotators for more accurate values. It then uses triangulation (Hartley and Zisserman, 2003) to obtain 3D joint coordinates. Therefore, the KSL-Guide has a much higher quality of data than PHOENIX14T, especially in subtle areas such as the fingers. We standardize and normalize the 3D joint coordinates to align all skeletons at the neck joints (Stoll et al., 2018) for both datasets.

### 4.2. Model Configuration

Our model builds on Pro-Transformer (Saunders et al., 2020b) and JoeyNMT (Kreutzer et al., 2019) using the PyTorch (Paszke et al., 2017) framework. The encoder and multi-channel spatio-temporal decoder are each configured with 2 layers, and the multi-head attention mechanism is equipped with 4 heads. The text/gloss embedding dimensionality is set to 512 in the experiments. In the multi-channel spatio-temporal decoder, we allocate embedding dimensions as follows: 256 for the face, 128 for both the right and left body poses, and 512 for the full-channel sign pose. Additionally, each attention layer incorporates a feed-forward dimension of 2048. To initialize our model, we employ the Xavier method (Glorot and Bengio, 2010). For parameter optimization, we leverage the Adam optimizer (Kingma and Ba, 2015) with a learning rate of  $10^{-3}$ . All models are trained with batch sizes of 64 and validated every 5000 steps.

### 4.3. Evaluation Metric

To evaluate the performance of the proposed model, we utilize a back-translation evaluation metric (Saunders et al., 2020b) that translates the produced sign pose sequences back to source sequences i.e. texts or glosses. For comparison with existing models, we utilize the commonly used SLT model (Camgoz et al., 2020). For further details on the model’s construction and training regimen, see (Saunders et al., 2021a). To quantify the performance on the PHOENIX14T dataset, we calculate the BLEU (n-grams ranging from 1 to 4) (Papineni et al., 2002) and ROUGE scores (Lin, 2004) for the back-translated text sequences. Higher BLEU

Model		Dev set					Test Set				
		BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE
Ground Truth		11.64	14.55	19.13	27.39	28.51	10.93	13.74	18.47	27.29	28.13
G2P	PT	10.12	12.60	16.64	24.74	26.16	9.41	12.09	16.64	25.20	25.77
	CasDual	11.20	14.02	18.50	26.93	28.58	11.16	13.89	18.50	26.70	28.02
	MCST-PA	11.28	14.13	18.82	27.75	<b>29.51</b>	10.60	13.53	18.34	27.18	27.93
	MCST-SO	10.79	13.65	18.32	27.05	28.91	10.90	13.77	18.53	27.27	28.16
	MCST-GF	<b>11.73</b>	<b>14.49</b>	<b>18.97</b>	<b>28.08</b>	29.39	<b>11.82</b>	<b>14.78</b>	<b>19.57</b>	<b>28.26</b>	<b>28.89</b>
T2P	PT	8.78	11.09	15.11	23.20	24.62	8.83	11.19	15.36	23.23	23.97
	CasDual	9.44	11.76	15.66	<b>23.58</b>	<b>25.09</b>	9.04	11.46	<b>15.62</b>	<b>23.77</b>	<b>24.72</b>
	MCST-PA	9.46	11.83	15.60	23.21	24.87	9.03	11.40	15.27	22.81	24.14
	MCST-SO	9.15	11.33	15.07	22.65	23.95	8.94	11.22	14.96	22.38	23.53
	MCST-GF	<b>9.65</b>	<b>12.07</b>	<b>15.97</b>	23.45	25.05	<b>9.21</b>	<b>11.53</b>	15.38	22.81	23.45

Table 1: Comparison of back-translation scores of Pro-Transformer (PT), CasDual-Transformer (CasDual), and MCST-Transformer on the PHOENIX14T Dataset. The table presents results for ground truth (our preprocessed) data, G2P and T2P tasks. MCST-Transformer models use abbreviations 'MCST-PA' for parallel addition fusion, 'MCST-SO' for serial operation, and 'MCST-GF' for gating fusion.

and ROUGE scores reflect improved model performance. Because the KSL-Guide dataset contains only glosses, we included word error rate (WER) (Ali and Renals, 2018) to indicate the model performance. For WER, lower scores indicate a better model performance.

#### 4.4. Tasks

We perform two different SLP tasks, namely, gloss to pose (G2P) and text to pose (T2P). G2P is the task of producing sign pose sequences from gloss sequences. T2P refers to the task of generating sign pose sequences from spoken word sequences.

## 5. Quantitative Results

In this section, we present the quantitative results for PHOENIX14T and KSL-Guide datasets. First, we compare the performance of our proposed model against existing approaches, achieving state-of-the-art results in both G2P and T2P tasks. Then, we conduct an ablation study on the PHOENIX14T dataset to assess the effectiveness of our introduced modules.

### 5.1. Baseline Comparison

We compare our model with Pro-Transformer and CasDual-Transformer. When training the Pro- and CasDual-Transformer, we maintain the original implementation reported in (Saunders et al., 2020b) and enhance them with gaussian noise and future predictions. We also implement these enhancement techniques in our MCST-Transformer. For a fair comparison, we use the same data (our preprocessed data) and model configurations to train all the models. Therefore, instead of using the results reported in (Saunders et al., 2020b), we reevaluate the back-translation scores of the baseline.

**PHOENIX14T Dataset.** As shown in Table 1, in all cases, our models employing three different fusion techniques, namely 'MCST-PA', 'MCST-SO', and 'MCST-GF', consistently outperform Pro-Transformer. The results support our claim that jointly learning channel-level features in both spatial and temporal dimensions can encourage SLP models to generate more expressive sign language poses. More specifically, our MCST-PA improves by 1.19 on the G2P task and 0.20 on the T2P task regarding the BLEU-4 score on the test set. MCST-SO showcases a more pronounced increase of 1.49 in the BLEU-4 score on the G2P task. However, when applied to the T2P task, the observed enhancement is comparatively limited, with a mere 0.11 increase in the BLEU-4 score. This result suggests that the sequential processing of temporal information before spatial information may not be the most effective approach for the T2P task. The top-performing MCST-GF model exhibits remarkable improvements, achieving a substantial increase of 25.6% (2.41 BLEU-4 score) on the G2P task and 4.3% (0.38 BLEU-4 score) on the T2P task on the test set compared to Pro-Transformer. Compared with CasDual-Transformer, MCST-GF improves 0.66 and 0.17 BLEU-4 scores on the G2P and T2P tasks, respectively. The results underscore the efficiency of implementing the gating fusion, which facilitates a dynamic balance between the spatial and temporal attention components.

**KSL-Guide Dataset.** Note that the BLEU and ROUGE scores obtained for this dataset were significantly higher than those obtained for the PHOENIX14T dataset. We hypothesize that this disparity arises because of the superior video resolution and abundance of training data in the KSL-Guide dataset, especially when considering its relatively smaller vocabulary size.

Specifically, MCST-PA outperforms the benchmarks, achieving an improvement of 1.52 in the BLEU-4 score and a reduction of 1.59 in the WER

Models	Dev set						Test Set					
	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE	WER	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE	WER
Ground Truth	96.52	96.93	97.24	97.81	97.00	2.92	94.96	95.43	95.97	96.56	96.38	4.16
PT	96.17	96.69	97.29	97.91	97.17	2.60	95.11	95.28	95.82	96.36	95.68	3.95
CasDual	97.51	97.54	97.74	98.12	97.52	2.29	<b>96.70</b>	96.89	96.98	97.29	97.05	2.91
MCST-PA	<b>97.58</b>	<b>97.96</b>	<b>98.26</b>	<b>98.64</b>	<b>98.55</b>	<b>1.66</b>	96.63	<b>97.02</b>	<b>97.36</b>	<b>98.02</b>	<b>97.59</b>	<b>2.39</b>
MCST-SO	96.07	96.46	96.84	97.50	96.90	3.02	96.40	96.56	96.72	97.29	96.81	3.23
MCST-GF	96.48	96.73	96.93	97.49	97.00	2.91	94.93	95.37	95.98	96.77	96.23	3.85

Table 2: Comparison of back-translation scores of Pro-Transformer (PT), CasDual-Transformer (CasDual), and MCST-Transformer on the KSL-Guide dataset.

Decoder Configuration		Dev set					Test Set				
		BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE
Ground Truth		11.64	14.55	19.13	27.39	28.51	10.93	13.74	18.47	27.29	28.13
G2P	PT	10.12	12.60	16.64	24.74	26.16	9.41	12.09	16.64	25.20	25.77
	MCTA	11.66	<b>14.57</b>	<b>19.22</b>	27.76	<b>29.64</b>	10.90	13.79	18.39	26.72	28.20
	MCSA	11.55	14.38	18.92	27.47	29.04	11.27	14.03	18.67	27.24	28.11
	MCST-GF	<b>11.73</b>	14.49	18.97	<b>28.08</b>	29.39	<b>11.82</b>	<b>14.78</b>	<b>19.57</b>	<b>28.26</b>	<b>28.89</b>
T2P	PT	8.78	11.09	15.11	23.20	24.62	8.83	11.19	15.36	<b>23.23</b>	23.97
	MCTA	8.95	11.15	14.94	22.67	24.34	8.37	10.72	14.68	22.10	22.96
	MCSA	9.47	11.83	15.74	<b>23.56</b>	<b>25.75</b>	8.80	11.05	14.89	22.40	23.62
	MCST-GF	<b>9.65</b>	<b>12.07</b>	<b>15.97</b>	23.45	25.05	<b>9.21</b>	<b>11.53</b>	<b>15.38</b>	22.81	<b>23.45</b>

Table 3: The back-translation scores of MCST-Transformer with different decoder configurations obtained on the PHOENIX14T dataset.

score compared to Pro-Transformer; a reduction of 0.52 of the WER score compared to CasDual-Transformer. Similarly, MCST-SO achieves an improvement of 1.29 in the BLEU-4 score and a decrease of 0.72 in the WER score compared with Pro-Transformer. It is observed that the parallel addition strategy outperforms the gate fusion approach on the KSL-Guide dataset, which, conversely, excels on the PHOENIX14T dataset. This discrepancy can be attributed to the relatively straightforward nature of the KSL-Guide dataset and the controlled environments in which it was gathered, which do not fully exploit the potential of the gate fusion mechanism. In contrast, the PHOENIX14T dataset, characterized by its rapid motion shifts and complex gestural expressions, effectively showcases the gating fusion’s capacity to mitigate noise and manage intricacies. Considering that sign language videos in real-life situations usually have poorer visual quality and more intricate gestural movements compared to those created in a studio setting, it is advisable to employ the gating fusion in real-world applications.

## 5.2. Ablation Study

To clarify the impact of each suggested component, we conduct ablation studies on both G2P and T2P tasks using the PHOENIX14T dataset. Table 3 presents the performance comparison of the proposed MCST decoder with various configurations. MCTA refers to the utilization of the multi-channel temporal attention module, whereas MCSA denotes the exclusive use of the multi-channel spatial attention module, with both models being immediately followed by a text-sign attention layer.

MCST-GF is our main contribution model, with the gating fusion.

From the results presented in Table 3, it is evident that for the G2P task, MCTA and MCSA yield comparable results. This underscores the efficacy of our spatial and temporal attention mechanisms even when deployed in isolation, indicating their potential to boost performance. The gating fusion offers a more flexible equilibrium between spatial and temporal attention attributes, leading to the creation of more articulate sign language sequences. For the T2P task, both MCTA and MCSA surpass Pro-Transformer (PT) in terms of performance on the validation set; however, they underperform on the test set. In contrast, the gating fusion rectifies this disparity and outperforms PT on the test set. One possible explanation for observing a varying performance regarding the T2P task could be the inherent differences between text and gloss. Unlike glosses, in which the sequence closely mirrors sign language, texts often differ significantly in word order from sign language expressions. This distinction may diminish the relevance of leveraging simultaneous spatial interactions, resulting in only marginal improvements in such scenarios.

## 6. Qualitative Analysis

This section provides a qualitative analysis of the produced sign poses on the PHOENIX14T dataset, as presented in Table 4. Specifically, we visualize the generated sign pose sequences of Pro-Transformer, CasDual-Transformer, and MCST-GF for the G2P task. It can be seen from the given example that Pro-Transformer leads to the generation



	FREITAG SONNE WOLKE SUEDE ANFANG REGEN						
Input	English Translation: Friday, sunshine and clouds mixed, showers early in the south.						
PT(Saunders et al., 2020b)							
CasDual(Ma et al., 2024)							
MCST-GF							
Ground Truth							
Original							
Frame #	1	2	3	4	5	6	7

Table 4: The qualitative results of G2P on the test set of PHOENIX14T dataset. The top row shows the input gloss and corresponding English translation. The second and third rows are the frames generated by Pro-Transformer and CasDual-Transformer, respectively. The fourth row is the produced frames by our best-performing MCST-GF model. The final two rows depict the ground truth sign language poses and their corresponding real-life frames.

of stagnant and repetitive sign poses, which are particularly noticeable in the 4th and 5th frames, as well as in the 6th and 7th frames. Compared to Pro-Transformer, CasDual-Transformer generates hand poses with richer details but less accurate hand positions. Our MCST-GF mitigates the issue of insufficient pose accuracy, especially with the improved synergy of multiple channels in these frames, but it still encounters challenges regarding gestures executed with a lesser range of motion and inadequately visible fine-grained actions. Two potential approaches can be considered to mitigate these issues: 1) integrating unsupervised training methods to expand the dataset, allowing the model to generate salient poses more effectively through extensive data training; 2) manually guiding the model to focus on salient poses by employing algorithmic approaches that learn salient poses from a few samples or by adding manual labels to salient frames.

## 7. Conclusions

We propose a Multi-Channel Spatio-Temporal Transformer (MCST-Transformer) model for sign language production, which aims to integrate both the spatial and temporal dimensions of sign language, leveraging their complex interplay. To achieve this, we incorporate spatial attention to perceive inter-channel correlations in the spatial dimension, and temporal attention captures the temporal dynamics and continuity of each channel. Furthermore, we explore three spatio-temporal fusion strategies: parallel addition, sequential operations, and gating fusion to enhance the capabilities of our model. Through extensive experiments conducted on both the PHOENIX14T and KSL-Guide datasets, we demonstrate the superiority of MCST-Transformer over existing approaches while also highlighting the specific contributions of each proposed module in improving sign language generation metrics.

## 8. Acknowledgements

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2024-RS-2023-00255968) grant, the ITRC (Information Technology Research Center) support program (IITP-2021-0-02051) funded by the Korea government (MSIT), and the Foreign Intelligence support program funded by Shijiazhuang Science and Technology Bureau (Project No. 20240024).

## References

- Ahmed Ali and Steve Renals. 2018. Word error rate estimation for speech recognition: e-wer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 20–24.
- J Andrew Bangham, SJ Cox, Ralph Elliott, John RW Glauert, Ian Marshall, Sanja Rankov, and Mark Wells. 2000. Virtual signing: Capture, animation, storage and transmission-an overview of the visicast project. In *IEE Seminar on speech and language processing for disabled and elderly people (Ref. No. 2000/025)*, pages 6–1. IET.
- Danielle Bragg, Oscar Koller, Mary Bellard, Laran Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, et al. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 16–31.
- Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793.
- Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. 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.
- Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Stephen Cox, Michael Lincoln, Judy Tryggvason, Melanie Nakisa, Mark Wells, Marcus Tutt, and Sanja Abbott. 2002. Tessa, a system to aid communication with deaf people. In *Proceedings of the fifth international ACM conference on Assistive technologies*, pages 205–212.
- Zhenchao Cui, Ziang Chen, Zhaoxin Li, and Zhaoqi Wang. 2022. Spatial-temporal graph transformer with sign mesh regression for skinned-based sign language production. *IEEE Access*, 10:127530–127539.
- Sarah Ebling and John Glauert. 2013. [Exploiting the full potential of jasigning to build an avatar signing train announcements.](#)
- Sarah Ebling and Matt Huenerfauth. 2015. Bridging the gap between sign language machine translation and sign language animation using sequence classification. In *Proceedings of SLPAT 2015: 6th workshop on speech and language processing for assistive technologies*, pages 2–9.
- Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and François Lefebvre-Albaret. 2012. The dicta-sign wiki: Enabling web communication for the deaf. In *International Conference on Computers for Handicapped Persons*, pages 205–212. Springer.
- Ralph Elliott, John RW Glauert, JR Kennaway, Ian Marshall, and Eva Safar. 2008. Linguistic modelling and language-processing technologies for avatar-based sign language presentation. *Universal Access in the Information Society*, 6(4):375–391.
- Sylvie Gibet, François Lefebvre-Albaret, Ludovic Hamon, Rémi Brun, and Ahmed Turki. 2016. Interactive editing in french sign language dedicated to virtual signers: Requirements and challenges. *Universal Access in the Information Society*, 15(4):525–539.
- John RW Glauert, Ralph Elliott, Stephen J Cox, Judy Tryggvason, and Mary Sheard. 2006. Vanessa—a system for communication between deaf and hearing people. *Technology and disability*, 18(4):207–216.
- Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings.

- Soomin Ham, Kibaek Park, YeongJun Jang, Youngtaek Oh, Seokmin Yun, Sukwon Yoon, Chang Jo Kim, Han-Mu Park, and In So Kweon. 2021. [KSL-Guide: A large-scale korean sign language dataset including interrogative sentences for guiding the deaf and hard-of-hearing](#). In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8.
- Richard Hartley and Andrew Zisserman. 2003. *Multiple view geometry in computer vision*. Cambridge university press.
- Wencan Huang, Wenwen Pan, Zhou Zhao, and Qi Tian. 2021. Towards fast and high-quality sign language production. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3172–3181.
- Eui Jun Hwang, Jung-Ho Kim, and Jong C Park. 2021. Non-autoregressive sign language production with gaussian space. In *Proceedings of the British Machine Vision Conference (BMVC)*.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*.
- Michael Kipp, Quan Nguyen, Alexis Heloir, and Silke Matthes. 2011. Assessing the deaf user perspective on sign language avatars. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, pages 107–114.
- Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. 2019. Joey nmt: A minimalist nmt toolkit for novices. 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): System Demonstrations*, pages 109–114.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Xiaohan Ma, Rize Jin, Jianming Wang, and Tae-Sun Chung. 2024. Attentional bias for hands: Cascade dual-decoder transformer for sign language production. *IET Computer Vision*.
- John McDonald, Rosalee Wolfe, Jerry Schnepf, Julie Hochgesang, Diana Gorman Jamrozik, Marie Stumbo, Larwan Berke, Melissa Bialek, and Farah Thomas. 2016. An automated technique for real-time production of lifelike animations of american sign language. *Universal Access in the Information Society*, 15(4):551–566.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.
- Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. 2020a. Adversarial training for multi-channel sign language production. In *Proceedings of the British Machine Vision Conference (BMVC)*.
- Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020b. Progressive transformers for end-to-end sign language production. In *European Conference on Computer Vision*, pages 687–705. Springer.
- Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021a. Continuous 3d multi-channel sign language production via progressive transformers and mixture density networks. *International journal of computer vision*, 129(7):2113–2135.
- Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021b. 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.
- Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2022a. 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.
- Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. 2022b. Skeletal graph self-attention: Embedding a skeleton inductive bias into sign language production. In *Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual: Challenges and Perspectives*, pages 95–102.
- Jr. Stokoe, William C. 2005. [Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf](#). *The Journal of Deaf Studies and Deaf Education*, 10(1):3–37.
- Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. 2018. Sign language production using neural machine translation and

generative adversarial networks. In *Proceedings of the British Machine Vision Conference (BMVC)*.

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):891–908.

Rachel Sutton-Spence and Bencie Woll. 1999. *The linguistics of British Sign Language: an introduction*. Cambridge University Press.

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.

World Health Organization. 2021. [World report on hearing - executive summary](#). Technical report.

Jan Zelinka and Jakub Kanis. 2020. Neural sign language synthesis: Words are our glosses. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3395–3403.

Inge Zwitterlood, Margriet Verlinden, Johan Ros, Sanny Van Der Schoot, and T Netherlands. 2004. Synthetic signing for the deaf: Esign. In *Proceedings of the conference and workshop on assistive technologies for vision and hearing impairment*. Citeseer.

## Appendices

### Appendix A. KSL-Guide qualitative results

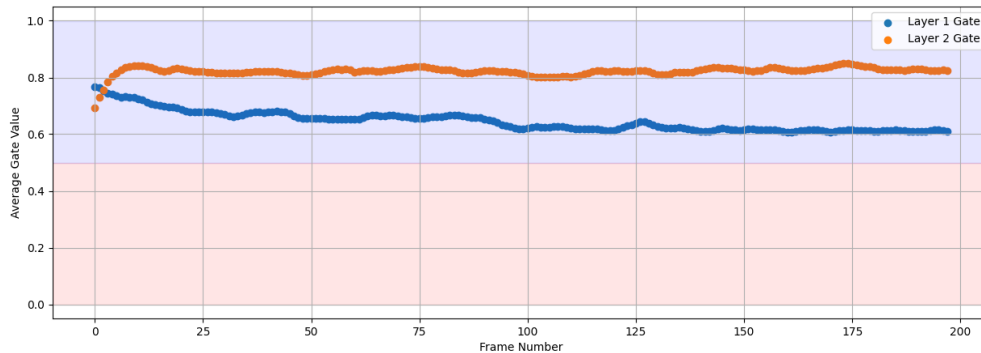
We also conduct a thorough analysis of the generated sign poses on the KSL-Guide dataset, identifying the areas in which our approach performs well and those that require further improvement. In Table 5, we can observe that, for the KSL-Guide dataset, both our MCST-Transformer and PT model produce considerably more accurate sign poses compared to their performance on the PHOENIX14T dataset. This observation is consistent with the elevated back-translation scores of the KSL-Guide dataset discussed in Subsection 5.1. Pro-Transformer seems ensnared by its tendency to regress to the mean, which occasionally results in less distinct or under-expressed poses, especially in Frames 2–6. Specifically, the misaligned hand orientations and aberrant hand shapes are apparent in these frames. CasDual-Transformer generates more accurate hand orientations but fails in the hand’s occlusion case in Frames 4 and 5. In contrast, our MCST-Transformer models generate vivid and articulated poses during these frames.

### Appendix B. Frame-level visualization of gate values

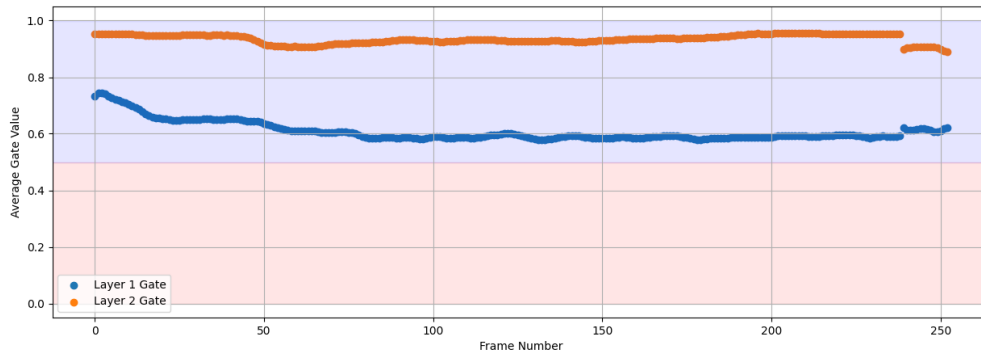
The visualization of the gate values in the gating

fusion is a critical component in understanding the weight distribution for the integration of spatial and temporal features during inference, as depicted in Figure 4. These gate values are bounded within the range of 0 to 1, where a value surpassing the midpoint of 0.5 signifies a predominant contribution from spatial features, whereas a value below this threshold indicates a more pronounced impact from temporal features, thus serving as a dynamic modulating factor for the balance of spatial and temporal attributes in the generated output.

We observe in the first gate layer an initial inclination towards spatial features in the early frames, suggesting an initial phase of learning focused on structural information. This spatial preference gradually transitions to a more balanced spatio-temporal contribution as the frames progress, reflecting the nuanced requirements of sign language sequence learning that necessitate an interplay of both spatial and temporal data. Conversely, the second gate layer exhibits a consistent preference for spatial features throughout. This dynamic interplay highlights the model’s ability to adaptively prioritize feature types according to the context provided by the sign language data.



(a) For the PHOENIX14T dataset



(b) For the KSL-Guide dataset

Figure 4: Visualization of frame-level gate values in sample sign language sequences.

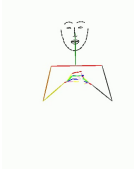

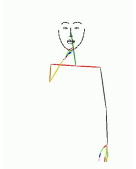
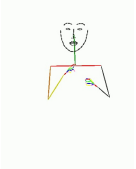
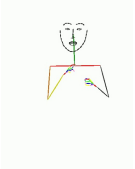
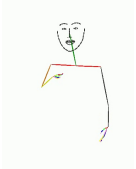
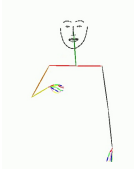
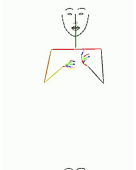


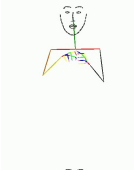
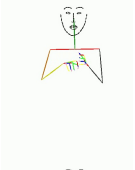
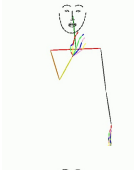
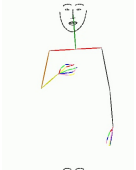
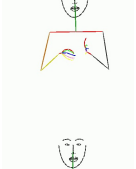
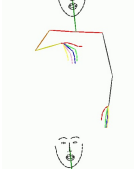
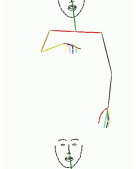
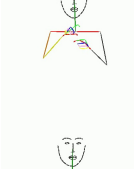
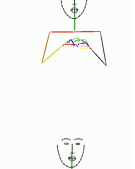
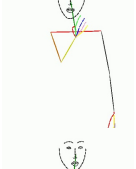
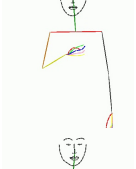
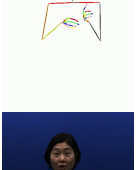
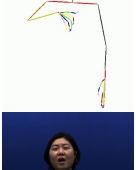
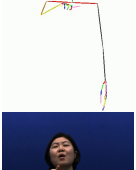
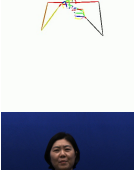
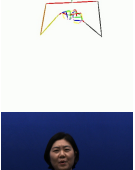
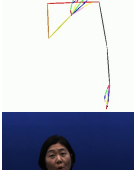
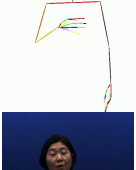
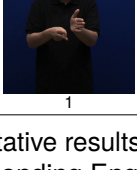
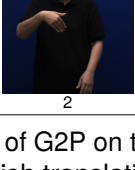
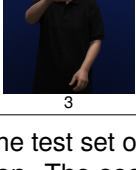

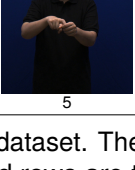
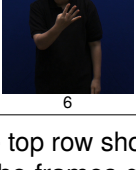
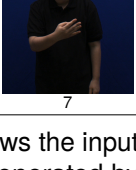
Input	서울역 가다 목적 지하철 번호 English Translation: The subway line number for the destination Seoul Station.						
PT(Saunders et al., 2020b)							
CasDual(Ma et al., 2024)							
MCST-GF							
Ground Truth							
Original Image							
Frame #	1	2	3	4	5	6	7

Table 5: The qualitative results of G2P on the test set of KSL-Guide dataset. The top row shows the input gloss and corresponding English translation. The second and third rows are the frames generated by Pro-Transformer and CasDual-Transformer, respectively. The fourth row is the produced frames by our best-performing MCST-PA model. The final two rows depict the ground truth sign language poses and their corresponding real-life frames.