

Improvement in Sign Language Translation Using Text CTC Alignment

Sihan Tan^{1,2}, Taro Miyazaki², Nabeela Khan¹, Kazuhiro Nakadai¹

¹Institute of Science Tokyo, ²NHK Science & Technology Research Laboratories
{tansihan, nabeela, nakadai}@ra.sc.e.titech.ac.jp
miyazaki.t-jw@nhk.or.jp

Abstract

Current sign language translation (SLT) approaches often rely on gloss-based supervision with Connectionist Temporal Classification (CTC), limiting their ability to handle non-monotonic alignments between sign language video and spoken text. In this work, we propose a novel method combining *Joint CTC/Attention* and *transfer learning*. The *Joint CTC/Attention* introduces hierarchical encoding and integrates CTC with the attention mechanism during decoding, effectively managing both monotonic and non-monotonic alignments. Meanwhile, *transfer learning* helps bridge the modality gap between vision and language in SLT. Experimental results on two widely adopted benchmarks, RWTH-PHOENIX-Weather 2014 T and CSL-Daily, show that our method achieves results comparable to state-of-the-art and outperforms the pure-attention baseline. Additionally, this work opens a new door for future research into *gloss-free* SLT using text-based CTC alignment.¹

1 Introduction

Sign language is a natural language used by the deaf and hard-of-hearing community for communication (Sutton-Spence and Woll, 1999), expressed through manual and non-manual elements, such as hand movements, fingerspelling, facial expressions, and body posture (Boyes Braem and Sutton-Spence, 2001). Due to the multi-modal nature of sign language, recent studies in sign language processing (SLP) aim to bridge the modality gap between vision and language. SLP encompasses several tasks, including sign language recognition (SLR) (Zuo et al., 2023; Sandoval-Castaneda et al., 2023), translation (SLT) (Camgoz et al., 2018; Zhang et al., 2023; Gong et al., 2024), and production (Rastgoo et al., 2021). In this paper, we

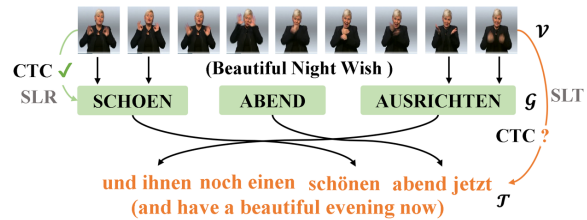


Figure 1: An overview of sign language recognition and sign language translation alignment. CTC is used for monotonic alignment between the sign language video \mathcal{V} and a gloss sequence \mathcal{G} . While the alignment between the sign language video with corresponding spoken text \mathcal{T} is non-monotonic.

focus on SLR and SLT. SLR involves identifying signs as glosses, a written representation of signs in the original order, whereas SLT focuses on translating the sign language video into its corresponding spoken text. Many SLR approaches utilize Connectionist Temporal Classification (CTC) loss (Graves et al., 2006) to localize and recognize glosses in an unsegmented sign language video (Cui et al., 2017; Min et al., 2021). However, in the context of SLT, CTC loss is typically employed only when glosses serve as an intermediate or auxiliary form of supervision (Camgoz et al., 2020; Yin and Read, 2020; Chen et al., 2022). This reflects the assumption that CTC’s monotonic alignment between input and output sequences is suitable for gloss-based supervision (Wong et al., 2024) but inadequate for the non-monotonic alignment required to directly translate sign language into corresponding spoken text, as displayed in Figure 1.

Nevertheless, many studies have successfully applied CTC to machine translation (MT) (Libovický and Helcl, 2018; Saharia et al., 2020) and speech translation (ST) (Chuang et al., 2021; Peng et al., 2024), demonstrating that CTC’s reordering capability holds promise for handling complex translation tasks. These approaches show that while CTC is effective for monotonic alignment tasks,

¹Code and resource are available: <https://github.com/Claire874/TextCTC-SLT>

it can be extended to non-monotonic alignment to address reordering challenges. This leads us to the question: *Can CTC be applied to non-monotonic alignment between signs and texts and benefit SLT?*

Inspired by the recent success of joint CTC/Attention in translation tasks (Yan et al., 2023), we investigate the potential of this approach for SLT. This combination provides a complementary solution: during encoding, hierarchical encoding effectively manages alignment (including sign language length adjustment and reordering), while joint decoding reduces exposure/label bias in the attention decoder by leveraging CTC’s conditionally independent alignment information. Our contributions are as follows:

- We propose leveraging a joint CTC/Attention framework combined with transfer learning to address the unique challenges of SLT, particularly the modality gap between visual sign language and spoken text. Compared with speech and text translation, SLT requires handling more complex spatial-temporal representations and addressing non-monotonic alignments to grasp the linguistic feature embedded in sign language.
- The proposed method achieves results comparable to state-of-the-art, outperforming the pure-attention baseline on widely adopted benchmarks, RWTH-PHOENIX-Weather 2014 T (Camgoz et al., 2018) and CSL-Daily (Zhou et al., 2021) (§ 5).
- We offer a promising direction for future research by exploring the potential of text CTC alignment for *gloss-free* SLT.

2 Background

Both CTC (Graves et al., 2006) and the attentional encoder-decoder paradigm (Vaswani et al., 2017) are designed to model Bayesian decision-making by selecting the output sequence \hat{Y} that maximizes the posterior likelihood $P(Y|X)$ from all possible sequences $\mathcal{V}^{\text{tgt}*}$. Here, $X = \{x_t \in \mathcal{S}^{\text{src}} | t = 1, \dots, T\}$ and $Y = \{y_l \in \mathcal{B}^{\text{tgt}} | l = 1, \dots, L\}$, where \mathcal{S}^{src} and \mathcal{B}^{tgt} denote the source and target language vocabularies, respectively, and T and L represent the lengths of the source and target sequences. The differences between the CTC and attention frameworks, a typical method of it is transformer-based encoder-decoder model, are illustrated in the following sections.

2.1 Comparison between CTC and Attention Framework

Hard vs. Soft Alignment. CTC marginalizes the likelihood over all possible input-to-alignment sequences $Z = \{z_t \in \mathcal{B}^{\text{tgt}} \cup \{\text{blank}\} | t = 1, \dots, T\}$, by introducing the blank label (Hannun, 2017). Each output unit z_t (whether blank or non-blank) is mapped to a single input unit x_t in a monotonic manner, resulting in hard alignment. In contrast, the attention framework allows a flexible input-to-output mapping via soft alignment, where an output unit y_l maps multiple input units $x_{[\dots]}$ or vice versa.

Conditional Independence vs. Conditional Dependence. CTC assumes conditional independence, meaning that each z_t is independent of $z_{1:t-1}$ if it is already conditioned on X (Nozaki and Komatsu, 2021). While the attention framework models each output unit y_l with conditional dependence on both the input X and previous output units $y_{l:l-1}$. This conditional dependence introduces potential label biases (Andor et al., 2016), where the model may over-rely on previous outputs during decoding.

Input-synchronous Emission vs. Autoregressive Generation. CTC is an input-synchronous model, producing an output unit at each input time step. As a result, CTC cannot generate an output sequence longer than the input, which simplifies the process by eliminating the need for explicit end detection. Conversely, the decoder in the attention framework is autoregressive, generating outputs one by one until a stop token $\langle \text{eos} \rangle$ is emitted.

As shown in (Yan et al., 2023), joint modeling of CTC/Attention during encoding and decoding benefits translation. During encoding, the hard alignment provided by CTC produces stable representations, enabling the decoder to learn soft alignment patterns more effectively. In joint decoding, the conditionally independent likelihood of CTC helps mitigate label biases in the conditionally dependent framework of attention. Additionally, the end detection issue in autoregressive generation is alleviated by the interaction between input-synchronous emission and autoregressive generation. *Will these advantages hold in SLT?*

2.2 Sign Language Translation

SLT can be viewed as a MT problem, with the distinction that the source language is represented by spatial-temporal pixels rather than discrete to-

kens (Chen et al., 2022). Intuitively, the joint CTC/Attention approach is well-suited for SLT, as it integrates the robust alignment capabilities of CTC with the contextual understanding provided by attention, effectively addressing the unique challenges of translating sign language into spoken text.

Our study focuses on translating a sign language video into a spoken text in an end-to-end manner, taking into account the information bottleneck that gloss introduces in cascading SLT, where gloss serves as the intermediate supervision (Yin and Read, 2020). Many end-to-end SLT approaches (Camgoz et al., 2020; Zhang et al., 2023; Ye et al., 2023; Tan et al., 2024) model the conditional dependent likelihood to generate spoken text. In general, given a sign video sequence $\mathcal{V} = \{v_1, v_2, \dots, v_{|\mathcal{V}|}\}$ consisting of $|\mathcal{V}|$ frames, the SLT model is designed to predict the ground truth reference $\mathcal{T} = \{t_1, t_2, \dots, t_{|\mathcal{T}|}\}$ which contains $|\mathcal{T}|$ words. A gloss sequence $\mathcal{G} = \{g_1, g_2, \dots, g_{|\mathcal{G}|}\}$ of length $|\mathcal{G}|$ employed as an additional regularization term. The core of these SLT approaches is a shared encoder (see Figure 3 left), trained with a joint loss to produce an encoded representation h , which is then used to predict both the gloss sequence \mathcal{G} and spoken text \mathcal{T} . For each time step l , there are:

$$P_{\text{CTC}}(g_l|\mathcal{V}) = \text{CTC}(h_l), \quad (1)$$

$$P_{\text{Attn}}(t_l|\mathcal{V}, t_{1:l-1}) = \text{Decoder}(h, t_{1:l-1}), \quad (2)$$

where $\text{CTC}(\cdot)$ refers to a projection onto the CTC output gloss $\mathcal{G} \cup \{\text{blank}\}$ followed by softmax. Meanwhile, $\text{Decoder}(\cdot)$ represents autoregressive decoder layers followed by a projection to the decoder output and softmax. For a given training sample consisting of a sign video, gloss sequence, and text, $(\mathcal{V}, \mathcal{G}, \mathcal{T})$, the whole SLT model is optimized using the maximum likelihood estimation (MLE) loss for SLT, and the CTC loss for SLR. The total training loss is formulated as:

$$\mathcal{L}(\mathcal{T}, \mathcal{G}|\mathcal{V}) = \mathcal{L}_{\text{MLE}}(\mathcal{T}|\mathcal{V}) + \alpha \mathcal{L}_{\text{CTC}}(\mathcal{G}|\mathcal{V}), \quad (3)$$

where hyperparameter α is utilized to balance the regularization effect of the CTC loss.

3 Model Architecture

As shown in Figure 2, our proposed method comprises three modules: sign embedding (SE), a hierarchical encoder conducting length adjustment and reordering, and a decoder that involves joint scoring during decoding. The training process is divided into two stages. Stage

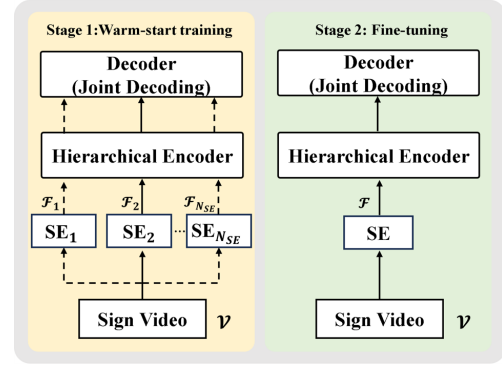


Figure 2: Overview of the proposed method: Stage 1: warm-start training using multiple sign embeddings with a hierarchical encoder and joint decoding. Stage 2: fine-tuning the model with a single sign embedding.

1: Warm-start training, where the parent model is trained using sign videos with multiple SEs ($SE_1, SE_2, \dots, SE_{N_{SE}}$). Each SE extracts different representations ($\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_{N_{SE}}$) from the sign video (\mathcal{V}), which are then passed through the hierarchical encoder and decoder. Stage 2: Fine-tuning the pre-trained model from Stage 1 with a single SE and child data pairs.

3.1 Sign Embedding Module

Since the sign video sequence is significantly longer than both the gloss and text sequences, *i.e.* $|\mathcal{V}| \gg |\mathcal{G}|$ and $|\mathcal{V}| \gg |\mathcal{T}|$, we down-sample the sign sequences to align them with the shorter gloss and text sequences. Depending on the training stage, we utilize either multiple or a single pre-trained visual model to extract frame-level representations from a sign video. The number of SEs (N_{SE}) is determined by dataset and model availability (§ 5.2). Following previous work, we use pre-trained SEs to extract sign representations from the video frames, which are then down-sampled through a linear layer. This process is formalized as:

$$\mathcal{F} = \text{SignEmbedding}(\mathcal{V}), \quad (4)$$

where $\mathcal{F} = \{f_1, f_2, \dots, f_{|\mathcal{F}|}\}$ denotes sign representations after the pre-trained SEs and subsequent down-sampling. Note that the parameters of the pre-trained SEs remain frozen throughout both training stages.

3.2 Hierarchical Encoding Module

Unlike the previous work on SLT adopting the shared encoder, we introduce a hierarchical encoder, which has been widely utilized in automatic speech recognition (ASR) (Krishna et al., 2018;

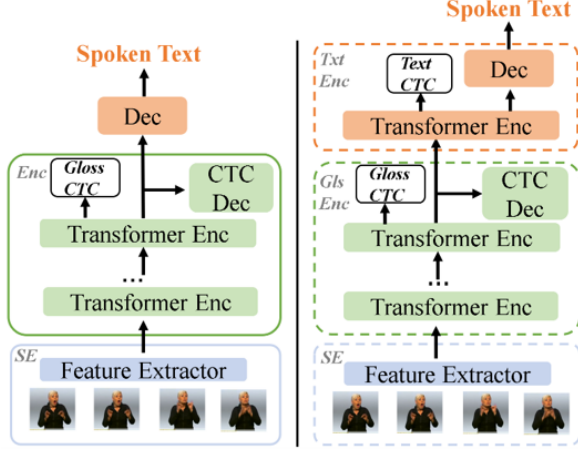


Figure 3: The shared SLT encoders (left): sign representations are down-sampled by a shared encoder using gloss CTC; hierarchical SLT encoders (right): sign representations are first down-sampled by $GlsEnc$ using gloss CTC ($GlsCTC$) and then reordered by $TxtEnc$ with text CTC ($TxtCTC$).

Higuchi et al., 2022). As shown in Figure 3 (right), our hierarchical encoder consists of gloss-oriented encoder ($GlsEnc$) and text-oriented encoder ($TxtEnc$) layers responsible for sign representations length adjustment and reordering, respectively.

Gloss-oriented Encoder. The $GlsEnc$ is trained using the CTC criterion, which aligns intermediate encoder representations with gloss sequences. The representations from the intermediate recognition encoder layers are then passed to the CTC decoder to generate gloss sequences.

Text-oriented Encoder. Next, $TxtEnc$ reorders the representations using hierarchical encoder layers. Here, the CTC criterion aligns the final encoder representations with spoken text sequences. While CTC is generally restricted to monotonic alignments, the neural network architecture enables latent reordering to handle the non-monotonic nature of text alignment. Our proposed SLT hierarchical encoder consists of the following components:

$$h^{gls} = GlsEnc(f), \quad (5)$$

$$P_{CTC}(g|f) = GlsCTC(h^{gls}), \quad (6)$$

$$h^{txt} = TxtEnc(h^{gls}), \quad (7)$$

$$P_{CTC}(z_t^{txt}|f) = TxtCTC(h^{txt}). \quad (8)$$

The $GlsEnc$ comprises N_{Gls} Transformer layers and a CTC loss calculation, while $TxtEnc$ consists of N_{Txt} Transformer layers and a CTC loss calculation.

We newly introduce $TxtCTC$ training to facilitate effective reordering in the encoding process. Using a multi-task objective, both hierarchical encoder and attentional decoder are jointly optimized:

$$\mathcal{L}_{SLT} = \lambda_1 \mathcal{L}_{GlsCTC} + \lambda_2 \mathcal{L}_{TxtCTC} + \lambda_3 \mathcal{L}_{MLE}, \quad (9)$$

where \mathcal{L}_{MLE} is a MLE loss for SLT, and λ_s decide the effects of gloss-oriented CTC, text-oriented CTC, and decoder MLE losses.

3.3 Joint Decoding Module

Our joint decoding adopts the *out-synchronous decoding* (i.e. attention plays a leading role) from (Yan et al., 2023) for its superior performance. However, ours differs in candidate selection: we first use the attention mechanism to propose candidates, then apply $TxtCTC$ constraints, offering greater flexibility and reducing label bias.

As shown in Algorithm 1, the joint decoding process has four steps: candidate selection, hypotheses expansion, joint scoring, and end detection. The algorithm generates candidates using the attentional decoder at each step l . A pre-defined beam size N_{beam} is adopted to select the top candidates (cand), starting with the partial hypotheses (hyp) from the previous step. Each hypothesis is extended by concatenating (\oplus) it with new candidates, adjusted by a dynamically computed length penalty β . The candidates are then scored using a joint mechanism P_{beam} combining normalized $TxtCTC$ scores ($\hat{\alpha}_{CTC}$) and Attention scores ($\hat{\alpha}_{Attn}$), where $TxtCTC$ (α_{CTC}) and Attention scores (α_{Attn}) represent their respective likelihoods. This joint scoring guides the selection of promising hypotheses. Updated partial hypotheses are stored in newHyp, while completed hypotheses ending with $\langle eos \rangle$ or reaching L_{max} are stored in finHyp.

3.4 Transfer Learning

One of the significant obstacles hindering the advancement of SLT is data scarcity (Moryossef et al., 2021; Zhu et al., 2023). To address this, many SLT studies mitigate the low-resource issue by transfer learning (Chen et al., 2022; Zhang et al., 2023). Transfer learning is widely used in low-resource scenarios to enhance model performance by leveraging knowledge from related tasks.

Warm-start Training. In transfer learning (Hadow et al., 2022), a parent model is pre-trained on a large corpus, and its parameters are used to initialize the child model on a smaller, target-specific

Algorithm 1 Joint Beam Search Decoding: The attentional decoder first proposes candidates for expanding hypotheses, ensuring all have l -length at step l .

```

1: procedure JOINT BEAM SEARCH DECODING( $hyp, h^{xt}, L_{max}, N_{beam}$ )
2:   newHyp={ }; finHyp={ }
3:   for  $t_{1:l-1} \in hyp$  do
4:     cand = top-k( $P_{attn}(t|h^{xt}, t_{1:l-1}), k=N_{beam}$ )
5:      $\beta = \text{Length\_Penalty}(l + 1)$ 
6:     cand = cand +  $\beta$ 
7:     for  $c \in cand$  do
8:        $t_{1:l} = t_{1:l-1} \oplus c$ 
9:        $\alpha_{CTC} = \text{TxtCTCScore}(t_{1:l})$ 
10:       $\alpha_{Attn} = \text{AttnScore}(t_{1:l})$ 
11:       $P_{Beam} = \hat{\alpha}_{CTC} + \hat{\alpha}_{Attn}$ 
12:      if ( $c$  is  $\langle eos \rangle$ ) or ( $l$  is  $L_{max}$ ) then
13:        finHyp [ $t_{1:l}$ ] =  $P_{Beam}(\cdot)$ 
14:      else
15:        newHyp [ $t_{1:l}$ ] =  $P_{Beam}(\cdot)$ 
16:      end if
17:    end for
18:  end for
19:  return newHyp, finHyp
20: end procedure

```

corpus. Warm-start training goes a step further by incorporating child language pairs into the parent model’s pre-training. After pre-training, the model is fine-tuned exclusively on the child language pairs. This approach benefits from the high likelihood of word or character overlap between those languages, improving the efficiency of transfer learning (Neubig and Hu, 2018). For this reason, we use parallel datasets to paraphrase spoken text by introducing normalization, lemmatization, and back translation. We adopt multiple pre-trained SEs for sign videos to extract varied sign representations from the same video.

Fine-tuning. The parent model is trained using sign representations from multiple SEs and combined spoken texts, while the child model is fine-tuned with a single SE and the original corpus.

4 Experimental Setup

Dataset. We evaluate the effectiveness of our proposed method on two popular benchmarks for SLT, RWTH-PHOENIX-Weather 2014 T (PHOENIX14T) (Camgoz et al., 2018) and CSL-Daily (Zhou et al., 2021). PHOENIX14T focuses on German sign language, sourced from weather broadcasting and CSL-Daily focuses on Chinese sign language, recorded in an ad hoc environment. Both datasets provide triplet samples: sign language video, a sentence-level gloss annotation, and

corresponding spoken text ($\mathcal{V}, \mathcal{G}, \mathcal{T}$).

Text Preprocessing. Following Zhu et al. (2023), we perform lemmatization and alphabet normalization on PHOENIX14T. Additionally, we apply back-translation using a text-to-gloss translation model to generate new source sentences from the glosses on the target side. For CSL-Daily, we follow a similar procedure, using a Zh-En model (Tiedemann and Thottingal, 2020) to generate parallel spoken text. The combined data is then used in warm-start training. Table 1 shows the statistics of the combined corpora.

Evaluation Metrics. We report results on SLT tasks using tokenized BLEU with n-grams from 1 to 4 (B@1-B@4) (Papineni et al., 2002) and Rouge-L F1 (ROUGE) (Lin, 2004). We employ word error rate (WER) to evaluate SLR performance.

Sign Embedding. Following recent studies (Camgoz et al., 2020; Zhou et al., 2021; Zhang et al., 2023) that use pre-trained models to extract sign video representations, we adopt multiple pre-trained visual models as SEs for warm-start training. For PHOENIX14T, we utilize pre-trained CNN layers from the CNN+LSTM+HMM setup (Koller et al., 2020), the pre-trained SMKD model² (Hao et al., 2021), and the pre-trained CorNet model³ (Hu et al., 2023). For CSL-Daily, we adopt the pre-trained SMKD model and CorNet to extract sign representations. The parameters of these pre-trained models are frozen throughout the training process. The child model utilizes a single SE, selected based on initial performance (§ 5.2).

Modeling. We compare our joint CTC/Attention model with purely attentional encoder-decoder baseline, Sign Language Transformer (Camgoz et al., 2020). The proposed and baseline models are trained separately with the same hyperparameters. Full details of the model sizes and hyperparameters are provided in the Appendix Table 6 and 7. Unless otherwise mentioned, we maintain these setups for the ensuing experiments.

5 Results and Analyses

We validate our proposed method on two popular benchmarks, PHOENIX14T and CSL-Daily, comparing it with both *cascading* and *end-to-end*

²https://github.com/VIPL-SLP/VAC_CSLR

³<https://github.com/hulianyuuy/CorrNet>

	PHOENIX14T						CSL-Daily					
	Text			Augmented text			Text			Augmented text		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
Sentences	7,096	519	642	21,288	519	642	184,401	1,077	1,176	36,802	1,077	1,176
Vocab.	2,887	951	1,001	3,770	951	1,001	2,343	1,358	1,358	2,491	1,358	1,358
#Words	99,081	6,820	7,816	308,198	6,820	7,816	291,048	17,304	19,288	362,778	17,304	19,288
#OOVs	–	57	60	–	55	58	–	64	69	–	43	47

Table 1: Statistics of the augmented corpora. "Sentences": the total number of sentences; "Vocab.": the number of spoken words (for Chinese, we count characters); "#Word": the total number of words; and "#OOV": the out-of-vocabulary words that appear in the development and test sets but are absent in the training set.

ID	Systems	PHOENIX14T	CSL-Daily
		B@4 ↑	B@4 ↑
1	Baseline	20.96	–
<i>Explore Sign Embeddings</i>			
2	1 + Replace sign embeddings to SMKD model	24.32	15.54
2.1	1 + Replace sign embeddings to Corrnnet model	24.64	13.02
2.2	1 + Combine all sign embeddings	24.38	13.77
<i>Explore CTC Capacity</i>			
<i>Hierarchical encoding</i>			
3	2 + <i>GlsCTC</i>	24.87	16.61
3.1	1 + <i>GlsCTC</i>	22.40	–
3.2	2.1 + <i>GlsCTC</i>	24.70	16.43
3.3	2.2 + <i>GlsCTC</i>	23.26	14.38
4	2 + <i>TxtCTC</i>	25.00	15.36
5	2 + <i>GlsCTC</i> and <i>TxtCTC</i>	25.84	18.27
<i>Joint decoding</i>			
6	4 + Joint decoding	25.63	15.79
7	5 + Joint decoding	26.37	19.17
<i>Explore Transfer Learning</i>			
8	3.2 + Warm start training	24.24	17.38
8.1	8 + Fine-tuning	25.26	18.01
9	7 + Combine all sign embeddings + warm start training	27.07	20.29
<i>Joint CTC/Attention</i>			
10	9 + Fine-tuning	27.93	22.04

Table 2: Ablation study of the proposed methods on dev sets. B@4: tokenized 4-gram BLEU.

state-of-the-art methods.⁴

5.1 Experimental Results

Our proposed method outperforms pure-attention encoder-decoder models on PHOENIX14T (see Table 3) and achieves results comparable to the state-of-the-art on CSL-Daily, as shown in Table 4. We conducted a series of ablation studies on their development sets, summarized in Table 2. On PHOENIX14T, it surpasses the baseline by 6.97 BLEU scores (Table 2: 1→10), demonstrating its effectiveness, particularly in handling non-monotonic alignment in SLT. Similarly, on CSL-Daily, consistent improvements with a gain of 6.50 BLEU scores are observed (Table 2: 2→10).

⁴The distinction between cascading and end-to-end is in § 2.2. Note that scores from previous papers are reported.

5.2 Exploring Sign Embeddings

As indicated by Zhang et al. (2023), high-quality SE benefits SLT, we explored the impact of different SEs on both SLR and SLT by replacing the baseline SE. Among all SEs, the pre-trained SMKD model shows superior performance in SLT on both datasets. Therefore, we used the SMKD model as the default SE in further analyses unless stated otherwise. We also examined SLR performances of all SEs. Interestingly, while combining all SEs slightly improves SLR performance on CSL-Daily, it underperforms in SLT compared to single embeddings such as SMKD or Corrnnet (see Table 5). Since both SMKD and Corrnnet are pre-trained using gloss annotations, their sign representations may share overlaps. In addition, glosses introduce information loss, limiting the benefit of combining

Methods	DEV		TEST				
	B@4	ROUGE	B@1	B@2	B@3	B@4	ROUGE
<i>Cascading</i>							
STMC-Transformer (Yin and Read, 2020)	24.68	48.70	50.63	38.36	30.58	25.40	48.78
ConSLT (Fu et al., 2023)	24.27	47.52	–	–	–	25.48	47.65
<i>End-to-end (Pure Attention)</i>							
Joint-SLRT (Camgoz et al., 2020)	22.38	–	46.61	33.73	26.19	21.32	–
Sign Back Translation (Zhou et al., 2021)	24.45	50.29	50.80	37.75	29.87	24.34	49.54
STMC-T (Zhou et al., 2022)	24.09	48.24	46.98	36.09	28.70	23.65	46.65
PET (Jin et al., 2022)	–	–	49.54	37.19	29.30	24.02	49.97
VL-Transfer (Chen et al., 2022)	27.61	53.10	53.97	41.75	33.84	28.39	52.65
SLTUNet (Zhang et al., 2023)	27.87	52.23	52.92	41.76	33.99	28.47	52.11
XmDA (Ye et al., 2023)	25.86	52.42	–	–	–	25.36	49.87
<i>Ours (Joint CTC/Attention)</i>	27.93	53.53	53.95	41.81	33.86	28.42	53.34

Table 3: Results of different systems on PHOENIX14T.

Methods	DEV		TEST				
	B@4	ROUGE	B@1	B@2	B@3	B@4	ROUGE
<i>Cascading</i>							
ConSLT (Fu et al., 2023)	14.80	41.46	–	–	–	14.53	40.98
<i>End-to-end (Pure Attention)</i>							
Joint-SLRT (Zhou et al., 2021)	11.88	27.06	37.38	24.36	16.55	11.79	36.74
Sign Back Translation (Zhou et al., 2021)	20.80	49.49	51.42	37.26	27.76	21.34	49.31
XmDA (Ye et al., 2023)	21.69	49.36	–	–	–	21.58	49.34
<i>Ours (Joint CTC/Attention)</i>	22.04	51.62	52.37	38.22	29.72	22.47	51.87

Table 4: Results of different systems on CSL-Daily.

Sign embedding (ID)	PHOENIX14T		CSL-Daily	
	WER↓	B@4↑	WER↓	B@4↑
Pre-trained CNN (3.1)	48.31	22.40	–	–
SMKD (3)	21.88	24.87	28.48	16.61
Cornnet (3.2)	21.43	24.70	30.54	16.43
Combining All (3.3)	23.01	23.26	28.16	14.38

Table 5: SLR performance of different sign embeddings.

these embeddings. However, as further discussed in § 5.4, transfer learning with fine-tuning can mitigate this issue and enhance performance.

5.3 Exploring CTC Capacity

TxtCTC plays a crucial role in both hierarchical encoding and joint decoding. It serves as a reordering tool for sign representations during encoding and provides the CTC score during decoding. To fully understand how each CTC component functions and verify whether they achieve the intended goals of length adjustment and reordering, we examined the respective contributions of *GlsCTC* and *TxtCTC*. First, we examined how both CTCs contribute to the model when only attention is used

in decoding (Table 2: explore hierarchical encoding). We then explored the contribution of *TxtCTC* in joint decoding by performing joint decoding without hierarchical encoding; in this case, only *TxtCTC* was involved in the whole process. Finally, we evaluated the performance of hierarchical encoding together with joint decoding.

The *TxtCTC* on its own appears to contribute more to PHOENIX14T than it does to CSL-Daily (Table 2: 3 and 4), indicating that reordering is more critical for German sign language. Hierarchical encoding yields more quality gains on the CSL-Daily dataset (Table 2: 2→5). Joint decoding with only *TxtCTC* achieves 25.63 BLEU and 15.79 BLEU on PHOENIX14T and CSL-Daily, respectively. This demonstrates the reordering capacity of CTC, especially when combined with attention mechanism for decoding, benefits SLT. Since no gloss is involved during training in ID 6 of Table 2, joint decoding with *TxtCTC* provides an alternative towards *gloss-free* SLT, which often struggles with non-monotonic alignments (e.g., Gong et al. (2024) present their *gloss-free* SLT achieves BLEU

score of 25.25 and 12.23 on development sets of PHOENIX14T and CSL-Daily). The combination of hierarchical encoding and joint decoding boosts SLT performance by 0.53 BLEU on PHOENIX14T and 0.90 on CSL-Daily (Table 2: 5→7).

5.4 Exploring Transfer Learning

In § 5.2, we demonstrated that simply combining all pre-trained SEs deteriorates SLT. We further investigated how this combination of SEs performs when paired with augmented spoken texts in transfer learning. For PHOENIX14T, we paired sign representations from pre-trained CNNs, CorNet, and SMKD with pre-processed, back-translated, and original spoken texts, respectively. For CSL-Daily, we paired sign representations from CorNet and SMKD with back-translated and original texts. As shown in Table 2, warm-start training mitigates the limitation of combining all SEs (+ 0.98 BLEU on PHEONIX14T, + 3.00 BLEU on CSL-Daily, Table 2: 3.3→8). While warm-start training enhances SLT by sharing mixed parameters, it lacks task-specific characteristics. To address this, we fine-tuned using SMKD sign representations and original spoken text. This further improves performance (Table 2: 8→8.1 and 9→10).

Besides, the case study of translation outputs is available in Appendix Table 8.

6 Related Work

The exploration of CTC’s latent alignment capacity has been widely conducted in speech translation (ST) and machine translation (MT) (Haviv et al., 2021). In ST, several studies have explored CTC’s potential to align and reorder spoken inputs with their target language translations. Chuang et al. (2021) found that CTC enhances the reordering behavior of non-autoregressive ST. Additionally, Zhang et al. (2022) demonstrated the effectiveness of CTC in ST by jointly training CTC and attention mechanisms without relying on ASR transcriptions. In MT, CTC has gained attention for both non-autoregressive (Saharia et al., 2020) and autoregressive paradigms (Yan et al., 2023).

Despite CTC’s reordering potential, research in SLT has underexplored this capability. In early SLT models, CTC was adopted primarily for its ability to align glosses with unsegmented sign video. Camgoz et al. (2020) pioneered SLT by jointly training CTC and attention mechanisms with glosses and corresponding spoken texts. Several subsequent

works (Zhou et al., 2021; Chen et al., 2022; Zhang et al., 2023; Ye et al., 2023; Tan et al., 2024) followed this paradigm, incorporating CTC to localize glosses and introduce techniques such as transfer learning and data augmentation to advance SLT. However, in these models, CTC’s role was limited to monotonic alignment, with little focus on its ability to reorder sequences. This approach, while functional, led to significant limitations in SLT. As noted by Yin et al. (2021), relying on glosses for SLT introduces an information bottleneck as the multi-dimensional sign video is converted into a linear gloss sequence. Additionally, glosses are linguistic tools not typically used by deaf and hard-of-hearing individuals for daily communication (Müller et al., 2023), further limiting their practical utility in SLT.

Given the limitations of glosses and the resource-intensive nature of gloss annotation (Uthus et al., 2023), recent research has shifted towards *gloss-free* SLT, which aims to translate sign video into text without gloss. This shift introduces challenges in aligning sign video with textual sequence. Wong et al. (2024) addressed this with a pseudo-gloss pretraining strategy in their Sign2GPT model for flexible alignment. Similarly, Gong et al. (2024) employed a Vector-Quantized visual sign module with a Codebook Reconstruction and Alignment module. Although *gloss-free* SLT has made significant progress, these approaches have moved away from using CTC, likely due to the assumption that CTC mainly supports monotonic alignment without fully exploring its reordering potential.

7 Conclusion

Unlike prior SLT approaches, this work revisits the reordering capability of CTC in the context of sign language translation (SLT). We propose a Joint CTC/Attention framework combined with transfer learning, leveraging CTC’s alignment and reordering capabilities for more effective SLT. Specifically, we introduce hierarchical encoding to manage length adjustment and reordering through the *GlsCTC* and *TxtCTC* training criteria, effectively addressing non-monotonic mappings in SLT. The joint decoding facilitates dynamic interaction between CTC and attention mechanisms. Additionally, transfer learning narrows the modality gap between vision (sign language) and language (spoken text). Our proposed method achieves translation results comparable to state-of-the-art. Moreover,

we demonstrate the strong potential of *TxtCTC* in joint decoding without gloss regularization, providing an alternative for *gloss-free* SLT to tackle non-monotonic alignments between sign language and spoken text.

8 Limitations

The limitations of our work lie in two points. First, data availability: beyond the inherent scarcity in sign language research, our approach relies on gloss and text annotations, available in only a few datasets. This limits the scalability and advancement of data-driven SLT models. Second, although our framework shows potential for *gloss-free* SLT, the current pre-trained sign embeddings still rely on gloss supervision, potentially introducing an information bottleneck, as discussed in § 5.2. In future work, we aim to explore sign representation strategies towards *gloss-free* SLT.

Acknowledgments

This work was partially supported by JST CREST JPMJCR19K1. The authors thank Yui Sudo for the insightful discussion and the anonymous reviewers for their valuable feedback.

References

- Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. [Globally normalized transition-based neural networks](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2442–2452, Berlin, Germany. Association for Computational Linguistics.
- Penny Boyes Braem and Rachel Sutton-Spence. 2001. *The Hands Are The Head of The Mouth. The Mouth as Articulator in Sign Languages*. Hamburg: Signum Press.
- 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 (CVPR)*, 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 (CVPR)*, pages 10023–10033.
- Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022. A simple multi-modality transfer learning baseline for sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5120–5130.
- Shun-Po Chuang, Yung-Sung Chuang, Chih-Chiang Chang, and Hung-yi Lee. 2021. [Investigating the re-ordering capability in CTC-based non-autoregressive end-to-end speech translation](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1068–1077, Online. Association for Computational Linguistics.
- Runpeng Cui, Hu Liu, and Changshui Zhang. 2017. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7361–7369.
- Biao Fu, Peigen Ye, Liang Zhang, Pei Yu, Cong Hu, Xiaodong Shi, and Yidong Chen. 2023. [A token-level contrastive framework for sign language translation](#). In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. 2024. Lfms are good sign language translators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18362–18372.
- Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. [Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks](#). In *International Conference on Machine Learning*, page 369–376. Association for Computing Machinery.
- Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. [Survey of Low-Resource Machine Translation](#). *Computational Linguistics*, 48(3):673–732.
- Awni Hannun. 2017. Sequence modeling with CTC. *Distill*, 2(11):e8.
- Aiming Hao, Yuecong Min, and Xilin Chen. 2021. [Self-mutual distillation learning for continuous sign language recognition](#). In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11283–11292.
- Adi Haviv, Lior Vassertail, and Omer Levy. 2021. [Can latent alignments improve autoregressive machine translation?](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2637–2641, Online. Association for Computational Linguistics.
- Yosuke Higuchi, Keita Karube, Tetsuji Ogawa, and Tetsunori Kobayashi. 2022. Hierarchical conditional

- end-to-end ASR with CTC and multi-granular subword units. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7797–7801. IEEE.
- Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. 2023. Continuous sign language recognition with correlation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2529–2539.
- Tao Jin, Zhou Zhao, Meng Zhang, and Xingshan Zeng. 2022. Prior knowledge and memory enriched transformer for sign language translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3766–3775.
- Oscar Koller, Necati Cihan Camgoz, Hermann Ney, and Richard Bowden. 2020. [Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(9):2306–2320.
- Kalpesh Krishna, Shubham Toshniwal, and Karen Livescu. 2018. Hierarchical multitask learning for CTC-based speech recognition. *arXiv preprint arXiv:1807.06234*.
- Jindřich Libovický and Jindřich Helcl. 2018. [End-to-end non-autoregressive neural machine translation with connectionist temporal classification](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3016–3021, Brussels, Belgium. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Yuecong Min, Aiming Hao, Xiujuan Chai, and Xilin Chen. 2021. Visual alignment constraint for continuous sign language recognition. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 11542–11551.
- Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. Data augmentation for sign language gloss translation. *arXiv preprint arXiv:2105.07476*.
- Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023. [Considerations for meaningful sign language machine translation based on glosses](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 682–693, Toronto, Canada. Association for Computational Linguistics.
- Graham Neubig and Junjie Hu. 2018. [Rapid adaptation of neural machine translation to new languages](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.
- Jumon Nozaki and Tatsuya Komatsu. 2021. Relaxing the conditional independence assumption of CTC-based ASR by conditioning on intermediate predictions. *arXiv preprint arXiv:2104.02724*.
- 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.
- Yifan Peng, Yui Sudo, Muhammad Shakeel, and Shinji Watanabe. 2024. [OWSM-CTC: An open encoder-only speech foundation model for speech recognition, translation, and language identification](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10192–10209, Bangkok, Thailand. Association for Computational Linguistics.
- Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, and Mohammad Sabokrou. 2021. Sign language production: A review. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 3451–3461.
- Chitwan Saharia, William Chan, Saurabh Saxena, and Mohammad Norouzi. 2020. [Non-autoregressive machine translation with latent alignments](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1098–1108, Online. Association for Computational Linguistics.
- Marcelo Sandoval-Castaneda, Yanhong Li, Diane Brentari, Karen Livescu, and Gregory Shakhnarovich. 2023. Self-supervised video transformers for isolated sign language recognition. *arXiv preprint arXiv:2309.02450*.
- Rachel Sutton-Spence and Bencie Woll. 1999. *The Linguistics of British Sign Language: An Introduction*. Cambridge University Press.
- Sihan Tan, Taro Miyazaki, Katsutoshi Itoyama, and Kazuhiro Nakadai. 2024. [SEDA: Simple and effective data augmentation for sign language understanding](#). In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 370–375, Torino, Italia. ELRA and ICCL.
- Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT—Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.
- Dave Uthus, Garrett Tanzer, and Manfred Georg. 2023. [YouTube-ASL: A large-scale, open-domain american sign language-english parallel corpus](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 29029–29047. Curran Associates, Inc.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz

- Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2024. [Sign2GPT: Leveraging large language models for gloss-free sign language translation](#). In *The Twelfth International Conference on Learning Representations*.
- Brian Yan, Siddharth Dalmia, Yosuke Higuchi, Graham Neubig, Florian Metze, Alan W Black, and Shinji Watanabe. 2023. [CTC alignments improve autoregressive translation](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1623–1639, Dubrovnik, Croatia. Association for Computational Linguistics.
- Jinhui Ye, Wenxiang Jiao, Xing Wang, Zhaopeng Tu, and Hui Xiong. 2023. [Cross-modality data augmentation for end-to-end sign language translation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13558–13571, Singapore. Association for Computational Linguistics.
- Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. [Including signed languages in natural language processing](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.
- Kayo Yin and Jesse Read. 2020. [Better sign language translation with STMC-transformer](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Biao Zhang, Barry Haddow, and Rico Sennrich. 2022. [Revisiting end-to-end speech-to-text translation from scratch](#). In *International Conference on Machine Learning*, pages 26193–26205. PMLR.
- Biao Zhang, Mathias Müller, and Rico Sennrich. 2023. [SLTUNET: A simple unified model for sign language translation](#). In *The Eleventh International Conference on Learning Representations*.
- Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. [Improving sign language translation with monolingual data by sign back-translation](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1325.
- Hao Zhou et al. 2022. [Spatial-temporal multi-cue network for sign language recognition and translation](#). *IEEE Transactions on Multimedia*, 24:768–779.
- Dele Zhu, Vera Czehmann, and Eleftherios Avramidis. 2023. [Neural machine translation methods for translating text to sign language glosses](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12523–12541, Toronto, Canada. Association for Computational Linguistics.
- Ronglai Zuo, Fangyun Wei, and Brian Mak. 2023. [Natural language-assisted sign language recognition](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14890–14900.

A Appendix

Hyperparameter	PHOENIX14T	CSL-Daily
N_{SE}	3	2
N_{Gls}	5	5
N_{Txt}	1	1
N_{Dec}	6	6
λ_1	1.0	5.0
λ_2	1.0	2.0
λ_3	1.0	1.0
L_{max}	30	50
Attention heads	4	4
Hidden size	256	256
Feed-forward dimension	4096	4096
Activation function	ReLU	Softsign
Learning rate	$1 \cdot 10^{-3}$	$1 \cdot 10^{-3}$
Adam β	(0.9,0.998)	(0.9,0.998)
Initializer	Xavier	Xavier
Init gain	0.5	0.5
Dropout rate	0.3	0.3
Label smoothing	0.1	0.1
Batch size	64	128
#params	28.95M	28.95M

Table 6: Hyperparameters and trainable parameters of our SLT models.

Decoding Type	Hyperparameter	Value
Pure attention	Penalty	0.6
	Beam size	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
	Penalty	0.6
Joint decoding	Beam size	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
	CTC score weight	0.3

Table 7: Decoding hyperparameters for pure attention and joint decoding methods.

Systems	Translation Output
<i>Examples from PHOENIX14T</i>	
Reference	dort morgen neunzehn im breisgau bis siebenundzwanzig grad (<i>there tomorrow nineteen in Breisgau up to twenty-seven degrees</i>)
Baseline	morgen neunzehn grad im breisgau bis siebenundzwanzig grad im breisgau (<i>tomorrow nineteen degrees in Breisgau up to twenty-seven degrees in Breisgau</i>)
Joint CTC/Attention	dort morgen bis neunzehn im breisgau bis siebenundzwanzig grad (<i>there tomorrow up to nineteen in Breisgau up to twenty-seven degrees</i>)
Reference	morgen regnet es vor allem noch in der südosthälfte (<i>tomorrow it will still be raining, especially in the southeast</i>)
Baseline	in the southeast it rains occasionally (<i>in the southeast it rains occasionally</i>)
Joint CTC/Attention	in der südosthälfte regnet es ergiebig (<i>heavy rain in the southeast</i>)
Reference	in der nacht ist es teils wolkig teils klar stellenweise gibt es schauer oder gewitter (<i>during the night it will be partly cloudy partly clear with showers or thunderstorms in places</i>)
Baseline	in der nacht ist es teils wolkig teils klar teils wolkig mit schauern und gewittern (<i>at night it will be partly cloudy partly clear partly cloudy with showers and thunderstorms</i>)
Joint CTC/Attention	in der nacht teils wolkig teils klar örtlich schauer oder gewitter (<i>at night partly cloudy partly clear local showers or thunderstorms</i>)
<i>Examples from CSL-Daily</i>	
Reference	蛋糕太甜，吃多了对身体不好。 (<i>Cake is too sweet and too much of it is not good for your health.</i>)
Baseline	蛋糕太甜，很安静。 (<i>The cake is too sweet and it's quiet.</i>)
Joint CTC/Attention	蛋糕太甜了，身体不健康 (<i>The cake is too sweet. It's not healthy.</i>)
Reference	我自愿加入组织。 (<i>I volunteer to join the organization.</i>)
Baseline	我自愿参加这项危险的任务。 (<i>I volunteer for this dangerous mission.</i>)
Joint CTC/Attention	我自愿参加的组织。 (<i>I volunteer for the organization.</i>)
Reference	他将来想成为科学家。 (<i>He wants to be a scientist in the future.</i>)
Baseline	他想做点兼职工作。 (<i>He wants to do some part-time work.</i>)
Joint CTC/Attention	他以后想做科学学校。 (<i>He wants to be a science school in the future.</i>)

Table 8: Case study of translation outputs on PHOENIX14T and CSL-Daily. Examples are from the development sets. Sentences in bracket are our English approximation translations. Our Joint CTC/Attention with transfer learning generally outperforms the baseline by providing translations that are more detailed, contextually accurate, and faithful to the original semantics of the source text.