CODA: A Novel End-to-End Computational Framework for Sign-to-Sign Translation Enhanced with LLMs

Anonymous ACL submission

Abstract

The number of different signed languages presents novel challenges in cross-cultural sign language processing. Our work takes a pioneering step into direct sign-to-sign translation across different sign language families. We first conduct a qualitative analysis of linguistic traits, both shared and distinctive, within a parallel corpus of multiple signed pairs of sentences. We then introduce a novel generation framework, CODA, for translating one sign language to another, employing Large Language models as intermediary text recognizers. We compile a dataset for sign-to-sign translation pairs across three signed languages: American Sign Language (ASL), Chinese Sign Language (CSL), and German Sign Language (DGS). We further utilize sign glosses as an intermediate representation to construct a multi-task model that can assist in preserving the semantic meaning of generated sign skeletal videos. We show that our model performs well on automatic metrics for sign-to-sign translation and generation as a novel first implementation. We make all our code and models available upon acceptance.

1 Introduction

Our paper tackles the computational challenges of translating between diverse sign languages, a crucial step toward enhancing accessibility and communication within deaf communities. The lack of a global standard for sign languages that fully accommodates the depth and complexity of regional sign languages poses significant barriers to cross-cultural communication.

Historically, distinct sign languages have developed across various regions since as early as the fifth century BC(Bauman, 2008), each with its own set of features and rules, from phonology and syntax to semantics and pragmatics(Virginia Swisher, 1988). These visual languages harness gestures, facial expressions, and the spatial dynamics of communication, leveraging shared cognitive abilities and linguistic conventions that, to some extent, unify signed languages globally. However, significant differences persist, highlighting the importance of technological interventions in bridging these gaps. For instance, despite English being a commonly spoken language in both the U.S. and the U.K., American Sign Language (ASL) and British Sign Language (BSL) remain mutually unintelligible(Pyers, 2012), underscoring the critical role of technology in exploring the meta-linguistic skills that transcend the spoken-sign language divide.

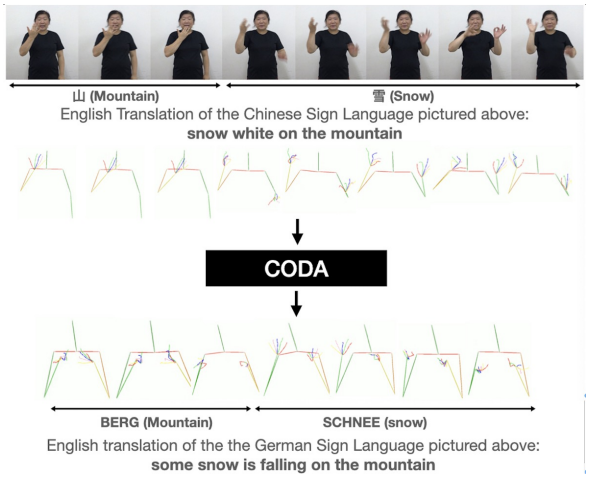

Figure 1: An example demonstrating how our CODA framework translates a sentence from one sign language to another.

As shown in Figure 2, the glosses with the same meaning (snowing) have shared spatial movements and are more understandable across different sign languages. Building on this inspiration and observation, this work aims to present the first study of a computational approach to automatically learn the mappings between sign languages and generate sign video translations across them.

Recent advancements in sign language technology have enhanced communication between sign and spoken language users. However, they fail to

address the critical communication issues between different groups of signers using separate signed languages. Inspired by the successes in spoken language Neural Machine Translation (Vaswani et al., 2017a; Liu et al., 2020; Xue et al., 2021), this study proposes translating sign videos across various sign languages to bridge these communication gaps.

Our work introduces a parallel corpus to enable communication between signers from different communities. Existing corpora, limited to single sign languages with spoken transcriptions and gloss annotations, lack cross-linguistic pairs with similar meanings, severely hindering the translation of sign language. We present an automatically aligned multilingual sign language corpus derived from multiple uni-language sign corpora to fill the gap of cross-lingual sign languages. It includes over 3,000 pairs covering ASL-CSL, DGS-ASL, and DGS-CSL parallel videos with the texts and glosses annotations. Text samples of our corpus are shown in Table 1. To our knowledge, this is the first corpus on the multilingual sign language dataset. Thus our effort not only facilitates cross-lingual sign language understanding but also offers insights into the social, cognitive, and linguistic nuances of sign languages, improving our comprehension of their use and processing across populations.

We begin by qualitatively analyzing the diversity and similarities in sign languages, offering insights into the challenges and opportunities in developing aligned sign language representations. With those insights, we introduce a framework for direct sign-to-sign translation, CODA. We propose an end2end model to translate sign videos from one language to another. We further investigated whether an auxiliary task for gloss detection could help the video translation quality. We further conducted experiments with Foundation Models such as GPT-4, in an attempt to refine the extraction of intermediate translations like glosses to enhance the capture of sign gloss patterns. Experiment results demonstrate that the proposed model generates moderate-quality videos and serves as a first step to mitigating the gap between different sign languages. We end the paper with suggestions for future research in the sign translation tasks.

2 Related Work

The study of sign languages has seen considerable advancements across two main fronts: sign language recognition (SLR) and sign language generation (SLG).

SLR has progressed from early visual recognition (Borg and Camilleri, 2019; Moryossef et al., 2020; Camgoz et al., 2018; Ko et al., 2019; Yin et al., 2021), segmentation (Fenlon et al., 2008; Cormier et al., 2016) efforts to sophisticated models capable of end-to-end translation (Starner, 1995; Yang and Sarkar, 2006; Huang et al., 2018; Camgoz et al., 2018), heavily relying on deep learning techniques like CNN/RNN(Huang et al., 2018; Cheng et al., 2020) and Transformer-based models (Yin and Read, 2020; Camgoz et al., 2020; Zhou et al., 2021b; Cheng et al., 2023; Wu et al., 2023) for state-of-the-art performances.

SLG, on the other hand, seeks to translate text into sign language poses or videos, where recent work has greatly benefited from larger datasets such as PHOENIX-14T (Camgoz et al., 2018) and CSL-Daily (Zhou et al., 2021a) and advanced neural networks, enabling the generation of accurate and expressive human skeletal sequences. (Stoll et al., 2018b; Zelinka and Kanis, 2020; Saunders et al., 2020a, 2021; Viegas et al., 2022; Zhou et al., 2021a)

Amidst these developments, translation and alignment in sign language translation have emerged as critical challenges. Notable efforts in this area include the application of neural machine translation (NMT) methods to translate spoken language text into sign language (SL) glosses (Zhu et al., 2023)(2023). They demonstrates substantial improvements on both German SL and American SL corpora. Similarly, earlier studies like Othman et al. (2011) and projects such as DeepASL Fang et al. have explored statistical and deep learning approaches to address the alignment and translation of English text to ASL gloss. Bidirectional translation systems, exemplified by Cate et al., have introduced generative models to enhance alignment between ASL and English, marking significant strides toward more nuanced translation mechanisms.

Our contribution diverges from these established paths by focusing on the direct translation between different sign languages, aiming to leverage the unique visual and linguistic features inherent to sign languages. This novel approach, which builds upon the foundational work in both SLR and SLG, as well as the specific translation and alignment challenges addressed by recent research, represents a pioneering effort to enable direct, meaningful communication across diverse sign language com-

Figure 2: The sign of the word snow/snowing in Chinese Sign Language (CSL) (Zhou et al., 2021a), German Sign Language (DGS) (Camgoz et al., 2018), and American Sign Language (ASL) (Duarte et al., 2021) with their glosses. Similar patterns of spatial movements (hands from top to bottom in the orange area) and hand gestures (bending fingers for symbols of snow flower, as shown in the blue area) are shared across three languages. On the other hand, the duration and repetition of hand movements differ (DGS repeats twice while signers in the other two languages put the hand down only once). Our study is not limited to single gloss detection but to video translation.

	Text	**Gloss**
ASL	Low self esteem or a low feeling about oneself is unfortunately very common.	N/A (not provided by the dataset)
CSL	自卑，就是人类常见的一种表现。 Inferiority is a common manifestation of human beings.	自卑<gloss> 人<gloss> 人<gloss> 见<gloss> 有 Inferiority <gloss> human <gloss> human <gloss> see <gloss> have
ASL	Good evening.	N/A
DGS	und damit schönen guten abend . and have a nice good evening.	BEGRUESSEN SCHOEN GUT ABEND BEGRUESSEN welcome good evening
CSL	山上雪白一片。 snow white on the mountain	山<gloss> 颜色<gloss> 雪白 mountain <gloss> color <gloss> snow white
DGS	an den bergen fällt etwas schnee . some snow is falling on the mountains.	BERG <gloss> SCHNEE mountain <gloss> snow

Table 1: Example of our constructed parallel corpora in texts, where in the real dataset, we have the video paired up as well. For non-English languages, we provide a Google translation for reference of meaning. Note that the ASL dataset is not released with glosses.

munities.

3 Qualitative Analysis of Sign Language

Sign languages can compress the information of similar spoken languages in different perspectives, which may vary in gesture movements, facial expressions, and duration of activity. We first analyzed the average length of sign video frames for glosses by dividing the frame counts by the gloss numbers of each instance. This is an approximation of the information compression for individual glosses. Since the How2Sign dataset (ASL) does not release the gloss, we use the texts instead. As shown in Table 2, ASL tends to utilize a longer time to present a single gloss, while DGS is the most efficient one. This may be affected by the domain

restriction of DGS where weather forecasting has a narrower vocabulary and thus is more elegant.

Linguistic Analysis

ASL	CSL	DGS
min/mean/max	min/mean/max	min/mean/max
0.1/7.9/115.0	1.6/17.3/73.4	3.2/15.5/71.5

Table 2: The average frame counts per gloss across the Sign Language Datasets.

Visual Modality Similarity Based on the visual modality, we hypothesize that the overlapping or similarity of sign languages can be attributed to the similarity of hand gestures used for specific signs. As the datasets studied in this paper belong to the continuous sign language domain and lack gloss-level mappings, we relied on an online sign language dictionary, SpreadTheSign[1], and conducted a qualitative analysis on thirteen selected triplets of signs across American Sign Language (ASL), Chinese Sign Language (CSL), and German Sign Language (DGS). When examining glosses associated with natural phenomena, such as "snow" and "mountains", we found that these three languages share similar gestures that mimic natural movements. However, when it comes to more abstract words like "sorry", the distinctiveness of spoken languages leads to significant differences in body language expressions. On the other hand, for words that involve measuring distance or length, such as "far" and "long", the three languages exhibit similarities by extending body regions. Overall, these findings highlight the effects of visual modality on sign languages.

4 Challenges in Cross-lingual Sign Translation

Although substantial efforts have been made in both Sign Language Recognition and Generation, connecting different sign languages proves to be challenging due to their distinctiveness. Glosses, which form the fundamental building blocks of sign language, can serve as a means of bridging the gap. One intuitive approach would be to employ a pipelined model to unify the two sections with natural languages (segmentation, then recognition of isolated natural language glosses from language one and generation based on the translations). However, certain challenges existed.[2]

Firstly, there is a lack of research on accurately segmenting and recognizing isolated signs in large-scale sign language datasets with an open vocabulary. The current state-of-the-art model(Renz et al., 2021) achieves an mF1B boundary prediction F1 score of only 0.53 on the widely used PHOENIX-14T dataset, limited to BSL and DGS. The largest word-level ASL dataset(Li et al., 2019) reports a Top-1 accuracy of 30% for a vocabulary of 2,000 words. Other sign languages have even smaller vocabularies, ranging from 40 (DGS)(Ong et al., 2012) to a few hundred (CSL), with low accuracy. These poor performances raise questions about the accuracy of isolated sign recognition for subsequent generation tasks.

Secondly, most works of continuous sign language generations have primarily focused on the PHOENIX-14T dataset, which comes from the narrowed domain of weather forecast and is only coupled with designed German sign glosses. This pattern also holds for other datasets, typically annotated individually and incorporating specialized glosses in their respective natural languages. Furthermore, these datasets often feature open vocabularies that differ significantly. There is a demand for a high-quality gloss-translation parallel corpus and a robust model that can generate sign videos from the given textual inputs to unify the language translation and later generations. Unfortunately, the sign generation results are still poor, given the automatic metrics, and not many explorations are done on other datasets.

5 Dataset Construction

In this section, We create a new Multilingual-SIGN corpus by pairing up sign videos from different sign language corpora. We have developed a meticulous matching methodology that considers the corresponding text transcriptions and gloss annotations associated with the sign videos. We describe the details below.

5.1 Curation of Raw Datasets

We obtain the raw dataset from the corresponding publications. In this task, we start with the three recently released continuous sign language datasets,

[1] https://www.spreadthesign.com/en.us/search/

[2] We experimented with a similar model that worked directly on the continuous signs in Appendix D and demonstrated the limitations of the pipelined attempt.

Dataset	Language	Samples (train/dev/test)	Data type	Sign Vocab
CSL-Daily	Chinese Sign Language	18,401 / 1,077 / 1,176	img;gloss;text	2,000
How2Sign	American Sign Language	31,128 / 1,741 / 2,322	video; img; text	16k
PHOENIX-14T	German Sign Language	7,096 / 519 / 642	video; gloss; text	1,066

Table 3: Continuous Sign Language Datasets.

CSL-Daily (Zhou et al., 2021a), How2Sign (Duarte et al., 2021), and PHOENIX-14T (Camgoz et al., 2018) – in which sign videos are cut into clips with individual sentences and their corresponding transcriptions.[3] Table 3 shows the statistics of sentences included in the three corpora.

5.2 Paraphrase Detection

We first aim to build a parallel corpus with cross-lingual sign language pairs. As the original corpora covered different data domains (i.e., CSL-Daily consists of daily life contents while PHOENIX-14T includes sign videos and corresponding text transcriptions from weather broadcasting series), estimating the degree of content overlap was challenging. Unlike traditional activity recognition tasks, directly classifying the long video clips as a single action is infeasible, as the video clips encoded sentences with complete meanings. Instead, we relied on the provided sentence transcriptions and gloss annotation to find the paraphrases across different datasets.

To tackle the issue that each sign language dataset has its spoken language, we first utilized a machine translation model[4] to translate all texts (Chinese and German) into English. Afterward, we relied on a neural paraphrase identification model (Reimers and Gurevych, 2019) to discover the paraphrases across the datasets. We carefully tune the threshold on the similarity scores with a held-out subset of human-annotated paraphrase pairs to guarantee the quality of extracted pairs. While it is possible that some pairs still lack a similar meaning, we considered the curated dataset as a valuable yet noisy training set for cross-lingual sign translations. Please refer to table 4 for detailed statistics on the final curated datasets. All three corpora except the How2Sign dataset have both gloss and text annotations. We obtained the predicted gloss from the state-of-the-art text-to-gloss translation models for the How2Sign dataset.

5.3 Construction and Postprocessing

Since multiple signers are signing the same sentence in the dataset, once we found the paraphrased sentence pair, we iteratively mapped the video pairs among the candidates' pool. This procedure dramatically enlarges the final dataset size but also introduces the issue of duplicated training signals. Yet, as the size of the sign language is orders of magnitude smaller than the spoken language machine translation corpus (i.e., several million pairs (Bojar et al., 2018)), we posit that such duplication can facilitate the model better to capture the nuance mapping between different sign languages. Once we obtained the video pairs, following prior work (Saunders et al., 2020b), we converted the sequence of sampled frame images into a 3D skeleton pose. The 2D skeletal joint positions are extracted from each video using OpenPose (Cao et al., 2019). We then lifted the 2D joints into 3D poses utilizing the skeletal model estimation improvements presented in (Zelinka and Kanis, 2020). Additionally, since those datasets are constructed with camera shots from different angles, we applied the skeleton normalization similar to (Stoll et al., 2018a). Regarding the text and glosses, for ASL and DGS, we do the normal space splitting. For CSL, we apply a Chinese text segmentation tool[5] on the texts for tokenization. We ended up with six pairs of parallel corpora.

Dataset	Train	Test
CSL-Daily - PHOENIX	2,274	669
How2Sign - PHOENIX	317	435
CSL-Daily - How2Sign	630	677

Table 4: Statistics of final constructed parallel dataset.

[3] For CSL-Daily, we have signed an agreement of data use and followed the regulations on the usage of dataset from http://home.ustc.edu.cn/~zhouh156/dataset/csl-daily/. The other two datasets are released publicly available for research purposes only, and we strictly followed the agreements.

[4] https://github.com/Helsinki-NLP/Opus-MT

[5] https://github.com/fxsjy/jieba

6 Model

In this section, we first introduce an end2end model that models the sign language video translation in an end-to-end manner (section §6.1). To better utilize the multi-modality of the sign languages, we further propose a model (Figure 3) that makes use of the corresponding glosses come with the input sign sequence to introduce more in-domain knowledge into the encoder part.

Language	$BLEU_1$	$BLEU_4$	ROUGE
Oracle ASL	18.90	2.93	17.51
Oracle DGS	30.42	12.36	30.10
Oracle CSL	23.30	2.25	23.19

Table 5: Sign Language Translation Results using the oracle skeletal joints, glosses and texts.

6.1 End2End Model

The main goal of the cross-lingual sign language translation model is to transform a signing video from the source language into the video in the target language. Formally, given a sign skeletal sequence $X = [x_1, ...x_N]$, a translation model aims to learn the conditional probability $p = (Y|X)$ where Y represents the corresponding language's skeletal pose coordinate sequence $Y = [y_1, ...y_T]$. We build a Transformer-based model (Vaswani et al., 2017b) as our baseline. This model can generate output skeletal sequence in an auto-regressive manner. Following prior work (Saunders et al., 2020b), we fed the encoded input skeletal joints sequence into a modified decoder, which employs a counter-based decoding mechanism to guide the generation of continuous joint sequences $y^{1:T}$ and to decide the end of the generated sequence. This strategy can be formulated as:

$$[\hat{y}^{t+1}, \hat{c}^{t+1}] = Model(\hat{y}^t|\hat{y}^{1:t-1}, x^{1:N}) \quad (1)$$

where $\hat{y}^{t+1}$ and $\hat{c}^{t+1}$ are the generated joint sequence and the counter value for the generated frame t+1. This generation model is trained using the mean square error (MSE) loss between the generated sequence $\hat{y}_{1:T}$ and the ground truth $y_{1:T}$ as $L_{MSE} = \frac{1}{T}\sum_{i=1}^{T}(y_i - \hat{y}_i)^2$.

6.2 End2End with an Auxiliary Task

We propose to frame the task as a multi-task problem and separate it into two subparts.
The first is source-side sign language recognition, where we use a continuous sequence-to-sequence learning function, CTC(Graves et al., 2006), for gloss recognition. Following prior work (Camgoz et al., 2020), given a video input V, we can obtain the gloss probabilities at each time stamp as $p(g_t|V)$, using a linear projection layer followed by a softmax activation function. We then utilize CTC to compute $p(G|V)$ by marginalizing over all possible Video to Gloss alignments as: $p(G|V) = \sum_{\pi \in B} p(\pi|V)$ where π is a path and B are the sets of all viable paths for the Gloss, as did in (Camgoz et al., 2020). The final recognition loss function is computed as $L_{Recog} = 1 - p(G^*|V)$ where G^* is the oracle path obtained from the dataset.

We optimize the recognition loss together with the aforementioned MSE loss for sign joint generation. The final loss is:

$$L = \alpha * L_{Recog} + L_{MSE}. \quad (2)$$

, where α is a tunable hyperparamter.

Language	$BLEU_1$	$BLEU_3$	$BLEU_4$
ASL > DGS	19.9	1.89	1.09
ASL > CSL	3.2	0.00	0.00
DGS > ASL	53.25	4.85	2.71
DGS > CSL	5.2	0.00	0.00
CSL > ASL	39.22	4.94	2.83
CSL > DGS	49.74	4.49	2.50

Table 6: Gloss translations using GPT-4 (source –> target) results on the test set.

6.3 Experiments using GPT-4

We also conducted experiments with Foundation Models, such as GPT-4, in an attempt to refine the extraction of intermediate translations.

6.3.1 Spoken language to Gloss

We first evaluate the accuracy of converting ASL text to ASL gloss, comparing manually annotated glosses from the How2Sign dataset with those generated by GPT-4. This comparison was crucial to assess the linguistic alignment and translation accuracy of GPT-4. Although GPT did pretty well in generating glosses, it was observed that manual glosses better captured the context and idiomatic expressions unique to ASL, a challenging aspect for AI models like GPT-4.

6.3.2 Gloss-to-Gloss

We further investigated the efficacy of converting one sign language gloss to another. Specifically, we

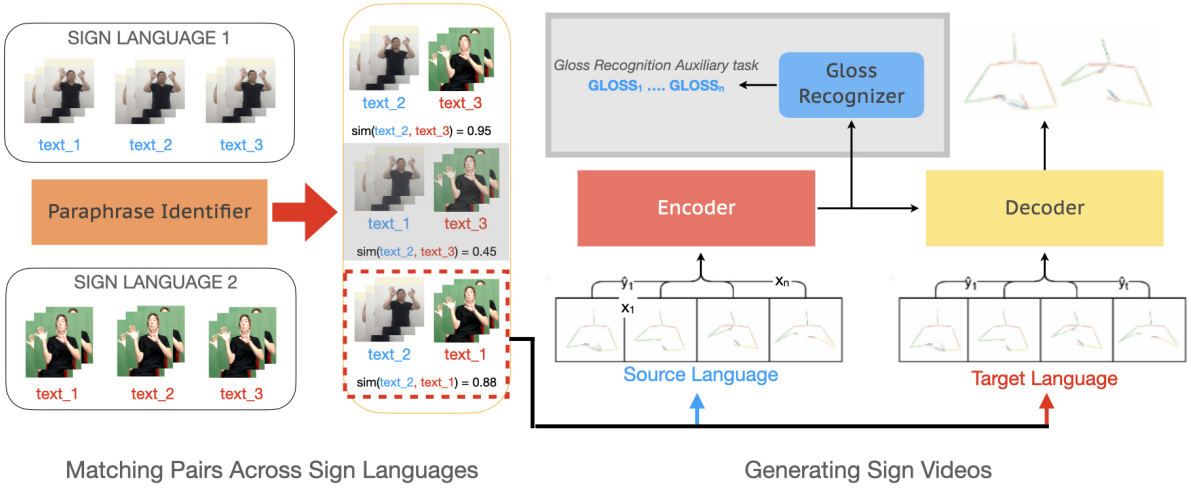

Figure 3: This figure shows the two stages of our CODA framework. We first include the identification and construction of parallel corpora using the transcriptions (§5).We then introduce an end2end model architecture with the additional gloss recognition auxiliary task (§6).

ASL –> X	CSL				DGS			
	$BLEU_1$	$BLEU_3$	$BLEU_4$	ROUGE	$BLEU_1$	$BLEU_3$	$BLEU_4$	ROUGE
end2end	17.00	0.94	0.00	16.46	14.18	6.80	5.73	13.22
end2end + recog	**17.16**	**1.19**	0.00	**16.82**	**15.86**	**8.08**	**6.81**	**14.53**
CSL –> X	ASL				DGS			
	$BLEU_1$	$BLEU_3$	$BLEU_4$	ROUGE	$BLEU_1$	$BLEU_3$	$BLEU_4$	ROUGE
end2end	**10.97**	**2.22**	**0.96**	10.32	**14.68**	**6.36**	**5.21**	13.91
end2end + recog	10.77	2.18	0.93	**10.34**	14.67	6.32	5.16	**14.09**
DGS –> X	ASL				CSL			
	$BLEU_1$	$BLEU_3$	$BLEU_4$	ROUGE	$BLEU_1$	$BLEU_3$	$BLEU_4$	ROUGE
end2end	9.01	**1.51**	**0.69**	7.71	**25.62**	0.00	0.00	**27.75**
end2end + recog	**9.29**	1.40	0.55	**7.90**	20.91	**4.34**	0.00	22.16

Table 7: Cross-lingual sign language pair translation (source –> target) results on the test set, **bolded** lines are better results of the two models.

focused on the translation between American Sign Language (ASL), Chinese Sign Language (CSL), and German Sign Language (DGS), exploring all possible translations of these glosses. The experiments were designed to understand the linguistic transformation and alignment challenges in sign language translation when done using foundation models.

We employed GPT-4 for gloss-to-gloss conversion, which was then evaluated against the ground truth of manually annotated glosses. This step required understanding the structural and idiomatic differences between languages.The results from this experiment can be observed in Table 6.

7 Evaluations and Results

We present the automatic metrics we use and the evaluation paradigm for signed languages in this section. Then we talk about our results of our experiments with these metrics.

7.1 Metrics and Back-Translation Model

To evaluate the generated skeletal joints' quality, following previous work (Saunders et al., 2020b; İnan et al., 2022), we back-translated the poses to the text domain and compared them with ground truth text, reporting ROUGE-L and BLEU scores for automatic evaluation. We provide the upper bound performances of the back-translation models built with SLT (Camgoz et al., 2020) in Table

5. Model implementation details are given in Appendix §B.

7.2 Automatic Results

We first train end-to-end baseline models on the six different language pairs. As shown in the first row of Table 7, the model performs best while translating ASL into the other two languages. These improvements could be attributed to the better back-translation quality than CSL and DGS (Table 5). At the same time, the ASL language is hard to back-translate, given its open vocabulary. This is amenable with recent studies on training sign language transformer model (Camgoz et al., 2020) over the How2Sign dataset (Duarte et al., 2022a). We also observe that although the model can translate high precision tokens from DGS to CSL and ASL, due to the narrow domain of the German Sign Language dataset (mainly weather forecasting), BLEU-4 scores are 0 for both models. With the introduction of the gloss recognition task, for ASL -> X tasks, we observe significant improvements across BLEU scores and ROUGE-L F1 scores. However, for CSL -> X tasks, the gloss does not help much. One other difference is that for DGS -> CSL tasks, though a lower $BLEU_1$ score is obtained with the introduction of the auxiliary task, we observe that the $BLEU_3$ score is improved. One of our main takeaways is that current advanced transformer-based models may not be able to generate satisfying results, especially given the noisiness of training and evaluation data. Meanwhile, when evaluating sign languages with a larger vocabulary and less repetitive patterns of inputs, current back-translation metrics fail to evaluate the quality of the generated videos. We leave this challenging problem to future work.

8 Discussions

We now discuss essential challenges that demand future efforts in sign-to-sign generation. One such challenge is the difficulties in cross-cultural alignments. Similar to spoken languages, sign languages can be affected by the physical and cultural factors of the user population. Thus, the representations of signs can be localized. Meanwhile, a lack of multilingual signers also hinders the iterations of model developments and evaluation. Current evaluations are restricted to the back-translation results of the generated sign videos, which lack spatial and temporal context, as discussed by (Inan et al., 2022). The lack of a proper evaluation metric remains a problem that needs to be addressed by an aggregated effort from different fields surrounding the sign language research community. Moreover, the fact that there are significantly few publicly available resources for sign language with glosses limited our choice and scope of datasets to the PHOENIX-14T and CSL-Daily dataset. The American Sign Language, such as How2sign (Duarte et al., 2021) came without oracle glosses, and we have to utilize non-perfect sign language translation models to derive glosses from the original text, thus introducing more errors.

9 Conclusions and Future Work

In this work, we address the problem of cross-lingual sign language translation, introducing a challenge for automatic video translation between sign languages. Our paper performs direct sign language translation on extracted human skeletal joints of videos to remove the overheads of pipelined Sign Language Generation (SLG) and Sign Language Recognition (SLR) tasks. We also release the first automatically aligned corpus with cross-lingual pairs that span three sign languages which can serve as a benchmark for future research. We demonstrate that incorporating the gloss information can assist in understanding the video, which highlights the need for using glosses to integrate more structure or stronger signals for better translation systems.

Future work could involve facial expression and motion capture for better understanding semantic meaning (Viegas et al., 2022) and spatial aspects of sign language. One other way is to employ the text-to-video retrieval approaches (Duarte et al., 2022b; Zuo et al., 2023) to verify the video alignments across different sign languages. Additionally, leveraging Large Language Models for refining dataset quality through better translation extraction and employing multimodal generative models(Li et al., 2022; Wu et al., 2023) could offer realistic sign video generation, aiding deaf community communication.

Ethics

We advocate for recognizing different signed languages and employing computational linguistics techniques for the preservation, documentation, understating, and generation of these languages. All models and analyses are built on publicly available

datasets. Privacy is an important issue in general in sign language processing. This work presents an example of ways that we can employ automatic skeleton and then avatar generation to preserve the singers' privacy. The generated human skeletal joints could be combined with an avatar and synthetic video techniques to create more real videos. Our work depends on pretrained models such as word and image embeddings. These models are known to reproduce and even magnify societal bias present in training data. Moreover, like many ML NLP methods, our methods are likely to perform better for content that is better represented in training, leading to further bias against marginalized groups. We can hope that general methods to mitigate harms from ML bias can address these issues.

Limitations

One limitation of our work is the cumulative error propagation that dissipates through the paraphrase identifier, sign language translation model, and back-translation, amplifying the total error. Due to the domain gap between different corpora, it is impractical to identify identical sign language video pairs based on transcriptions for those with longer and more complicated meanings. Experimental results demonstrate the need for better-constructed large-scale datasets with high-quality alignments and a more focused study from the linguistics perspective. Though imperfect, we hope this work could stimulate future studies looking at the cross-lingual aspects of sign languages and assist prospective sign language users in communicating better and breaking the language barrier.

Another limitation of this work is on the evaluation side. The current back-translation method is the only tool we have for evaluation on the text side, and we have limited resources available in the sign language area. Meanwhile, directly evaluating signs differs greatly from gesture evaluation, while gestures seem particularly vaguer than signs, and the accuracy and evaluation required are challenging. We encourage researchers from CV and NLP areas to work more closely and bridge the gap of understanding sign languages across multimodalities.

References

Dirksen Bauman. 2008. Open your eyes: Deaf studies talking. *University of Minnesota Press*.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Mark Borg and Kenneth P. Camilleri. 2019. Sign language detection "in the wild" with recurrent neural networks. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1637–1641.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *2018 IEEE/CVF 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*.

Hardie Cate et al. 2017. Bidirectional american sign language to english translation. In *Proceedings of the 2017 Conference on Sign Language Translation*.

Ka Leong Cheng, Zhaoyang Yang, Qifeng Chen, and Yu-Wing Tai. 2020. Fully convolutional networks for continuous sign language recognition. In *European Conference on Computer Vision*, pages 697–714. Springer.

Yiting Cheng, Fangyun Wei, Bao Jianmin, Dong Chen, and Wen Qiang Zhang. 2023. Cico: Domain-aware sign language retrieval via cross-lingual contrastive learning. In *CVPR*.

Kearsy Cormier, Onno Crasborn, and Richard Bank. 2016. Digging into signs: Emerging annotation standards for sign language corpora.

Amanda Duarte, Samuel Albanie, Xavier Giró-i Nieto, and Gül Varol. 2022a. Sign language video retrieval with free-form textual queries. *arXiv preprint arXiv:2201.02495*.

Amanda Duarte, Samuel Albanie, Xavier Giro i Nieto, and Gul Varol. 2022b. Sign language video retrieval with free-form textual queries. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro i Nieto. 2021. How2sign: A

large-scale multimodal dataset for continuous american sign language.

Biyi Fang et al. 2018. Deepasl: Enabling ubiquitous and non-intrusive word and sentence-level sign language translation. In *Proceedings of the 2018 Conference on Sign Language Translation Technology*.

Jordan Fenlon, Tanya Denmark, Ruth Campbell, and Bencie Woll. 2008. Seeing sentence boundaries. *Sign Language Linguistics*, 10:177–200.

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*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256. PMLR.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. 2018. Video-based sign language recognition without temporal segmentation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Mert İnan, Yang Zhong, Sabit Hassan, Lorna Quandt, and Malihe Alikhani. 2022. Modeling intensification for sign language generation: A computational approach. *arXiv preprint arXiv:2203.09679*.

Mert Inan, Yang Zhong, Sabit Hassan, Lorna Quandt, and Malihe Alikhani. 2022. Modeling intensification for sign language generation: A computational approach. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2897–2911, Dublin, Ireland. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2015. Adam: a method for stochastic optimization. In *3rd International Conference on Learning Representations*.

Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. 2019. Neural sign language translation based on human keypoint estimation. *Applied Sciences*, 9(13):2683.

Dongxu Li, Cristian Rodriguez-Opazo, Xin Yu, and Hongdong Li. 2019. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1448–1458.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Amit Moryossef, Ioannis Tsochantaridis, Roee Aharoni, Sarah Ebling, and Srini Narayanan. 2020. Real-time sign language detection using human pose estimation. In *European Conference on Computer Vision*, pages 237–248. Springer.

Eng-Jon Ong, Helen Cooper, Nicolas Pugeault, and R. Bowden. 2012. Sign language recognition using sequential pattern trees. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2200–2207.

Achraf Othman and Mohamed Jemni. 2012. English-asl gloss parallel corpus 2012: Aslg-pc12.

Achraf Othman et al. 2011. Statistical sign language machine translation: from english written text to american sign language gloss. In *Proceedings of the 2011 Workshop on Sign Language Translation*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.

J.E. Pyers. 2012. Sign languages. In V.S. Ramachandran, editor, *Encyclopedia of Human Behavior (Second Edition)*, second edition edition, pages 425–434. Academic Press, San Diego.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Katrin Renz, Nicolaj C. Stache, Neil Fox, Gül Varol, and Samuel Albanie. 2021. Sign segmentation with changepoint-modulated pseudo-labelling. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3398–3407.

Ben Saunders, Necati Cihan Camgöz, and R. Bowden. 2020a. Adversarial training for multi-channel sign language production. *ArXiv*, abs/2008.12405.

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. 2021. Mixed signals: Sign language production via a mixture of motion primitives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1919–1929.

Thad Starner. 1995. Visual recognition of american sign language using hidden markov models.

Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and R. Bowden. 2018a. Sign language production using neural machine translation and generative adversarial networks. In *BMVC*.

Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. 2018b. Sign language production using neural machine translation and generative adversarial networks. In *29th British Machine Vision Conference (BMVC 2018)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Carla Viegas, Mert İnan, Lorna Quandt, and Malihe Alikhani. 2022. Including facial expressions in contextual embeddings for sign language generation.

M. Virginia Swisher. 1988. Similarities and Differences between Spoken Languages and Natural Sign Languages. *Applied Linguistics*, 9(4):343–356.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Ruiduo Yang and S. Sarkar. 2006. Detecting coarticulation in sign language using conditional random fields. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 2, pages 108–112.

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.

Jan Zelinka and Jakub Kanis. 2020. Neural sign language synthesis: Words are our glosses. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3384–3392.

Hao Zhou, Wen gang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021a. Improving sign language translation with monolingual data by sign back-translation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1325.

Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. 2021b. Spatial-temporal multi-cue network for sign language recognition and translation. *IEEE Transactions on Multimedia*.

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 *CVPR*.

A ASL gloss extraction

We retrained a sign language translation model which produces glosses from the texts using the transformer-based model (Yin and Read, 2020). The model is trained on ASLG-PC12 (Othman and Jemni, 2012), which contains 87,709 training pairs. Following the setup in (Yin et al., 2021), we used their pre-processed glosses as the target.

B Model Implementation Details

We implemented all models for the sign video translation task based on the codebase released by (Saunders et al., 2020b). Different from their gloss/text to sign language generation, we modified the encoder part to accept human skeletal joints as inputs. For the end2end model, Both the encoder and decoder are built with two layers, 4 heads and embedding size of 512. We apply Gaussian noise with a noise rate of 5, as proposed by Saunders et al. (2020b). All network parts are trained with Xavier initialization (Glorot and Bengio, 2010), Adam optimization (Kingma and Ba, 2015) with default parameters and a learning rate of 1e-3. The model takes 3 hours

to train on 1 NVIDIA RTX 5000 GPU. We keep the model size fixed for our proposed model with auxiliary tasks. The output layer for gloss recognition has a dimension of 512. The model takes 4 hours to train on 1 NVIDIA RTX 5000 GPU. For the end2end model, we search the recognition loss weight α between (1, 0.1, and 0.01), and use 0.01 in the final result table.

We implemented the back-translation model on top of the original SLT code (Camgoz et al., 2020). The transformer models are built with one layer, two heads, and an embedding size of 128. The feature size is changed to 150, which is the sequence length of generated skeleton joints sequence. The recognition loss weight and translation loss weight are set to 5 and 1 for CSL and DGS back-translation models. We set the recognition loss of 0 for ASL, given that the dataset does not come with oracle gloss annotation. Back-translation models take around 1-3 hours for training and evaluation for all three languages. All models introduced above are implemented with Pytorch (Paszke et al., 2019).

C Error Analysis

We present a qualitative error analysis on Table8.

D Pipe-lined Model

For the pipelined model, we build the pipelines as follow: for each source sign language, we reuse the back-translation model that can recognize texts from the continuous skeletal joints sequneces. For machine translation, we use Google Translate to translating the recognized texts into the corresponding language. We further feed the translated results into the corresponding Progressive-Transformer based models (Saunders et al., 2020b) that are trained on the 3 datasets. For ASL, we find that the first stage recognizer performed poorly and failed to recognize the accurate meanings of ASL videos. We thus experimented with the pair of DGS-CSL, where the models are working relatively better. We reported the result of DGS-CSL translation: BLEU-1 15.75, BLEU-2 of 1.10, BLEU-4 of 0.0 and ROUGE-L of 16.57, which is worse than end2end models (bottom right corner) in Table 7 (BLEU-1 25.62 and ROUGE of 27.75).

To go beyond the limitations of automatic back-translation metrics and investigate how our system generates the videos, we perform a qualitative analysis of our model outputs, both on back-translated texts (Table 8). One issue is the low BLEU_4 score of 0 for the ASL/DGS to CSL translation task. As shown in the first row of Table 6, since the Chinese texts are pre-tokenized with the tokenization tool, it is less usual that continuous 4-grams appear in both the reference and oracle texts. Meanwhile, for ASL and CSL datasets with open-domain vocabularies, current alignments are not perfect enough and may introduce errors in the training stage. For instance, in the second example, there is no mention of specific food names for breakfast in the source video of ASL. However, both the generated video and the automatically paired reference video in CSL surprisingly produced *milk* as one of the foods ordered/eaten. This can be related to the domain of CSL, which covers entities that appear much in our daily life. Meanwhile, for ASL and CSL to DGS generation tasks we could look at the back-translated results to examine the generation quality. As illustrated in the third row of Table 8, though over-generating the "good evening" spans, the back-translation result matches the paired DGS target sentences. However, several pairs in the test set have distinct meanings, as shown in the last row. Such noises in the dataset can misguide the model and make the generated results nonsense.

Source Text	**Generated** (back-translation)	**Paired Target**
da haben wir am morgen **schnee** und schneeregen . Here we have **snow** and sleet in the morning.	明天‖ 白天‖ 会‖ 下雨 。 It will rain during the day tomorrow.	下雪 ‖ 了‖，‖ 今天‖ 真冷。 It's **snowing**, it's so cold today.
Now, a typical day starts with **breakfast**.	我‖ 想‖ 一杯‖ 牛奶 ‖，‖ 你‖ 要‖ 什么‖ 饮料‖？ I want a glass of *milk*, what do you want?	早饭‖ 我‖ 吃‖ 的‖ 是‖ 面包 ‖ 和‖ 牛奶‖ 。 I had bread and *milk* for **breakfast**.
Hi	hallo und guten abend **Hello** and good night	hallo und guten abend **Hello** and good night

Table 8: Qualitative analysis of model outputs, for non-English texts, we provide English translations (underlined) in the bottom of each row. **Bold** words are correctly translated across languages. For Chinese texts, we use the symbol "‖" to mark the tokenized word boundaries of prediction, which leads to the poor $BLEU_4$ performance in Table 7. We find that, although sometimes the automatically aligned pairs do not covey the identical meaning, our model can produce reasonable results and covering salient tokens. The examples are selected from DGS-CSL, ASL-CSL, and ASL-DGS from top to bottom.