How to *Align* Multiple Signed Language Corpora for Better *Sign-to-Sign* Translations?

Mert İnan[1], **Yang Zhong**[2], **Malihe Alikhani**[1]
[1] Khoury College of Computer Sciences, Northeastern University, Boston, USA
[2] School of Computing and Information, University of Pittsburgh, Pittsburgh, USA
inan.m@northeastern.edu, yaz118@pitt.edu, m.alikhani@northeastern.edu

Abstract

With over 300 documented signed languages worldwide, they serve as essential gateways for researchers and educators to explore cross-cultural and cross-linguistic influences on automatic sign recognition and generation. However, research in this field remains severely limited by scarce resources, particularly when analyzing multiple signed languages at once. In this work, we hypothesize that a linguistically informed alignment algorithm can improve the results of sign-to-sign translation models. To this end, we first conduct a qualitative analysis of similarities and differences across three signed languages: American Sign Language (ASL), Chinese Sign Language (CSL), and German Sign Language (DGS). We then introduce a novel generation and alignment algorithm for translating one sign language to another, exploring Large Language Models (LLMs) as intermediary translators and paraphrasers. We also compile a dataset of sign-to-sign translation pairs between these languages. Our model trained on this dataset performs well on automatic metrics for sign-to-sign translation and generation. Our code and data will be available for the camera-ready version of the paper. Our code and data are available at: https://github.com/Merterm/sign2sign

1 Introduction

Despite the growing need for advanced signing technologies, signed language (SL) resources remain scarce, posing significant challenges for computational linguistic research and accessibility within Deaf or Hard-of-Hearing (DHH) communities. A crucial step toward improving these technologies is the development of larger, multilingual signed corpora. While recent efforts have focused on collecting pre-aligned video datasets—such as (Yin et al., 2022) curating videos for 10 SLs and (Gueuwou et al., 2023) compiling Bible-based videos in 47 SLs—the real challenge lies in the automatic alignment of signs rather than just data collection.

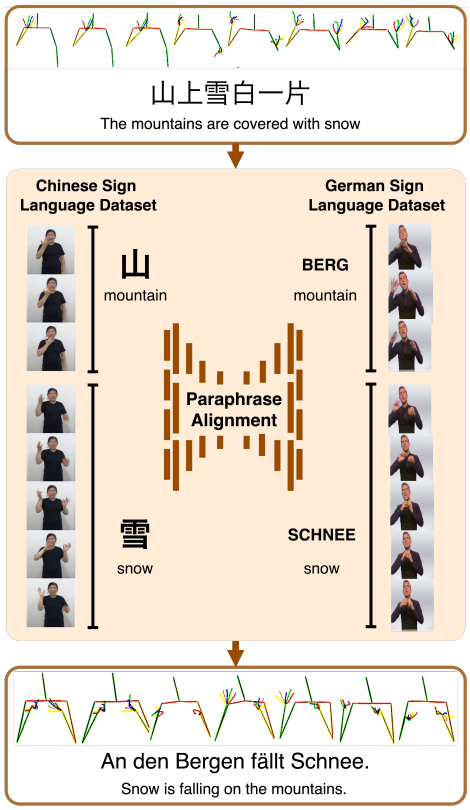

Figure 1: This figure depicts how we translate from one sign language to another. Inputs are shown at the top (skeletal body positions for CSL), and then the outputs are at the bottom for DGS. Glosses, which are intermediary textual representations (such as, '山' for the sign for 'mountain' in CSL), are also used with sign sequences. At its core, a paraphrasing alignment algorithm matches instances between different corpora.

Even though these are significant data collection initiatives, the main bottleneck lies in automatically *aligning* signs instead of collecting *pre-aligned* videos. So, in this paper, we work towards ways to align sign language representations automatically across three signed languages: American Sign Language (ASL), Chinese Sign Language (CSL), and German Sign Language (DGS).

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4003–4016

	Text	Gloss
ASL	Low self esteem or a low feeling about oneself is unfortunately very common.	N/A
CSL	自卑，就是人类常见的一种表现。 *Inferiority is a common manifestation of human beings.*	< 自卑 > < 人 > < 人 > < 见 > < 有 > <*inferiority*> <*human*> <*human*> <*see*> <*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 welcome*
CSL	山上雪白一片。 *The mountains are covered with snow*	< 山 > < 颜色 > < 雪白 > <*mountain*> <*color*> <*snow white*>
DGS	An den Bergen fällt Schnee. *Snow is falling on the mountains.*	BERG SCHNEE *mountain snow*

Table 1: This table shows examples of our constructed parallel corpus with ASL-CSL, ASL-DGS, and CSL-DGS translations. We use our alignment algorithm to get these final results. Here, glosses are intermediary textual representations for signs in the given source and target SLs. As the How2Sign dataset does not contain original glosses, we do not show them here. English translations are given underneath the original text for interpretability.

We claim that ways of automatically aligning can benefit from recent Neural Machine Translation (NMT) techniques, as well as LLM advancements (Vaswani et al., 2017a; Liu et al., 2020; Xue et al., 2021). These recent advancements in sign language translation models have enhanced communication between sign and spoken language users. However, they fail to address the critical communication issues between different groups of signers who use separate SLs. To ameliorate, we study ways to combine the power of LLMs, cognitive science, and linguistic findings to create more effective intermediary translators or paraphrasers between SLs. In this paper, we first present a linguistic analysis of the differences and similarities of SLs in a data-driven manner. Then, we propose a paraphrase alignment algorithm as shown in Figure 1. Using this approach, we present a multilingual SL corpus derived from multiple uni-language datasets. It includes over 3,000 pairs covering ASL-CSL, DGS-ASL, and DGS-CSL parallel videos with textual annotations (Table 1 shows our corpus's text samples). To our knowledge, this is the first corpus that automatically aligns multiple corpora for a unified multilingual SL dataset. Finally, we train a generation model using this dataset and report experimental results on its performance in sign-to-sign translation tasks. We hope that this dataset and the automatic alignment algorithm will enable future research on the communication between signers from different communities.

2 Related Work

The study of signed 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.

Concurrently, SLG has greatly benefited from larger datasets, such as PHOENIX-14T (Camgoz et al., 2018) and CSL-Daily (Zhou et al., 2021a), gradually increasing the accuracy of the generations (Stoll et al., 2018b; Zelinka and Kanis, 2020; Saunders et al., 2020a, 2021; Zhou et al., 2021a; Moryossef et al., 2021; Lin et al., 2023; Müller et al., 2023b). There have also been works focusing on more prosodic generations —i.e., the intensity, duration, and repetition of signs — with better use of the signing space and with more awareness of the intensifiers in facial expressions (Inan et al., 2022; Viegas et al., 2023).

Figure 2: The sign of the word 'snow' 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 are shared across these SLs in terms of the use of signing space (e.g., hands moving from top to bottom, as shown in the orange area), and hand gestures (e.g., bent fingers as an iconic representation of a snowflake, as shown in the blue areas) are shared across three languages. Contrarily, the duration, and repetition of hand movements may differ due to other linguistic factors (e.g., the sign is repeated twice in DGS, while only once in CSL and ASL).

Amidst these developments, translation and alignment in SLG have emerged as critical challenges. Notable efforts in this area include applying NMT methods to translate spoken language text into SL glosses (Zhu et al., 2023). They demonstrate substantial improvements in both DGS and ASL corpora. Similarly, earlier studies like Othman et al. (2011) and projects such as DeepASL (Fang et al., 2018) have explored statistical and deep learning approaches to address the alignment of English text and ASL gloss. Bidirectional translation systems, exemplified by Cate et al. (2017), have introduced generative models to enhance alignment between ASL and English, marking significant strides toward more nuanced translation mechanisms. Also, established challenges (Müller et al., 2023a) and review articles (De Coster et al., 2023) for translations between spoken and signed languages–such as Swiss German and Swiss German Sign Language (DSGS)–have shown the successes and challenges of current state-of-the-art neural translation systems.

In addition to bilingual translation systems, there has been recent interest in collecting and curating multilingual sign translation corpora. Notably, works of (Yin et al., 2022) curating a subset of videos from the SpreadTheSign initiative to form a dataset with parallel data for 10 SLs across various domains, and (Gueuwou et al., 2023) presenting a single-domain signed videos in 47 SLs.

Our contribution builds on these established paths by focusing on the in-the-wild alignment between different SLs–where already aligned signed videos are not present–aiming to leverage unique linguistic features inherent to SLs. 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 communities.

3 Aligning Different Signed Languages

To align different SL corpora, we first analyze them linguistically. In a data-driven manner, we also identify the challenges of aligning these separate datasets. As a result of these analyses, we propose an approach that uses a paraphrase detection module.

Qualitative Linguistic Analysis Historically, distinct SLs have evolved across various regions since as early as the 5th 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 lan-

guages harness gestures, facial expressions, and the signing space, leveraging shared cognitive abilities and linguistic conventions that, to some extent, unify SLs globally. However, significant linguistic differences exist among SLs and between them and spoken languages, 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 incomprehensible (Pyers, 2012), as the former has been developed from Old French Sign Language, various village SLs, and home sign systems; in contrast, the latter is developed from Old British Sign Language. These differences underscore the critical role of technology in exploring the meta-linguistic skills that transcend the spoken-sign language divide.

As shown in Figure 2, the signs with the same meaning (snowing) have shared signing spaces and are more interpretable across different SLs. Building on this observation, we hypothesize that signed data (containing videos, body positions, text, and glosses) that has already been collected for different SLs under different corpora can be aligned.

Quantitative Linguistic Analysis Signed languages can represent information in different aspects, such as gestures, movements, facial expressions, and sign duration. Different SLs also have different prosodic characteristics. To quantify these differences, we first analyze the average length of signed video frames by dividing the frame counts by the number of signs in the video. This can be regarded as an approximation of the information representation statistic for an individual sign. We use glosses as proxies to detect the number of signs in the video. Since the How2Sign dataset (ASL) does not contain glosses, we use the texts instead. We show these preliminary analyses in Table 2.

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 sign across different signed languages based on How2Sign (ASL), CSL-Daily (CSL), and PHOENIX-14-T (DGS) datasets.

It can be observed that ASL tends to utilize a longer time to present a single sign, while DGS uses a shorter duration on average. Even though this is a loose approximation of the information content per sign, it provides an initial understanding of how the temporality of signed languages may affect the alignment. Based on these analyses, we further discuss the challenges of sign-to-sign translation in Appendix A.

4 Our Approach

In this section, we introduce our approach to address the challenges of sign-to-sign translation. We focus on combining already-present sign video datasets in any language and align with another dataset using paraphrase detection algorithms. With this approach, we create a new corpus by pairing sign videos from CSL-Daily, PHOENIX-14T, and How2Sign corpora. In this section, we describe the details of our alignment methodology, inspired by our lexical analyses in Section §3. This happens in three steps: preprocessing, paraphrase detection, and postprocessing.

4.1 Preprocessing: Curation of Raw Datasets

In this task, we start with three well-established continuous SL 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.[1] Each of these datasets already contains its preprocessing steps, and they are standardized across different corpora using methods such as video clipping, signer cropping, and masking. We aggregate all of the sentences, their corresponding video clips, and their corresponding glosses (if glosses are unavailable, we obtain the predicted gloss from the state-of-the-art text-to-gloss translation models.) in a single format across these datasets. Table 3 shows the statistics of the three corpora.

4.2 Paraphrase Alignment

To align these different datasets, we focus on the core NLP task of paraphrase detection applied to SLs. As the original corpora cover different data domains (i.e., CSL-Daily consists of daily life contents while PHOENIX-14T includes sign videos and corresponding text transcriptions from

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

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: This table presents several statistics about the continuous SL datasets. These are some of the most commonly used corpora in the sign language processing literature. Each data collection effort has its own shortcomings and advantages, e.g. How2Sign contains more than 10k samples, but this makes it difficult to manually annotate signs with intermediary textual representations, i.e. glosses.

weather broadcasting series), estimating the degree of content overlap is challenging. Unlike traditional activity recognition tasks, directly classifying long video clips as a single action is infeasible, as the video clips encode sentences with complete meanings. Instead, we focus on "finding the paraphrases" across different datasets using the provided textual representations (i.e., sentence transcriptions and gloss annotations).

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. Each number corresponds to the number of samples of signed sentences using the given SL dataset pair.

We first utilize an open-source machine translation model[2] to translate all texts (Chinese and German) into English. Afterward, we employ a neural paraphrase identification model (Reimers and Gurevych, 2019) to identify the paraphrases across all these datasets. We 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 consider the curated dataset as a valuable yet noisy training set for cross-lingual sign translations. Table 4 presents the statistics on the final curated datasets.

4.3 Construction and Postprocessing

After identifying candidate sentence pairs using the paraphrase alignment strategy, we perform multiple stages of postprocessing to construct our dataset. After alignment, we end up with multiple candidate videos and sentence pairs corresponding to the same segment–this is due to multiple signers in each dataset signing the same segment. In these instances of multiple candidates, we map every source sign to all the different target signs as separate samples. This procedure dramatically enlarges the final dataset size but also introduces the issue of duplicated training signals. Yet, as the size of any SL corpus is orders of magnitude smaller than a multilingual 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 nuanced mapping between different SLs.

Once we obtain the video pairs, following prior work (Saunders et al., 2020b), we convert the sequence of sampled frame images into 3D body poses with skeletal coordinates. To accomplish this, we first extract 2D skeletal joint positions—i.e., coordinates corresponding to the location of upper-body joints (hands, arms, torso) in the two-dimensional space (please refer to Figures 1 or 3 for visual representations of these skeletal body positions—from each video using OpenPose (Cao et al., 2019). As the next and final step, these 2D coordinates are converted to three dimensions by utilizing skeletal model estimation techniques as presented in (Zelinka and Kanis, 2020). Additionally, we apply body pose coordinate normalizations similar to (Stoll et al., 2018a), to account for skewed coordinates as the original datasets are constructed with different camera angles. Finally, to post-process the text and glosses for ASL and DGS data, we split the tokens using whitespace. For CSL, we apply a Chinese text segmentation tool[3] on the texts for tokenization. After these stages, we end up with our multimodal and multilingual dataset for three parallel SLs (i.e., ASL-CSL, ASL-DGS, CSL-DGS). We then use this dataset to train a translation generation model.

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

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

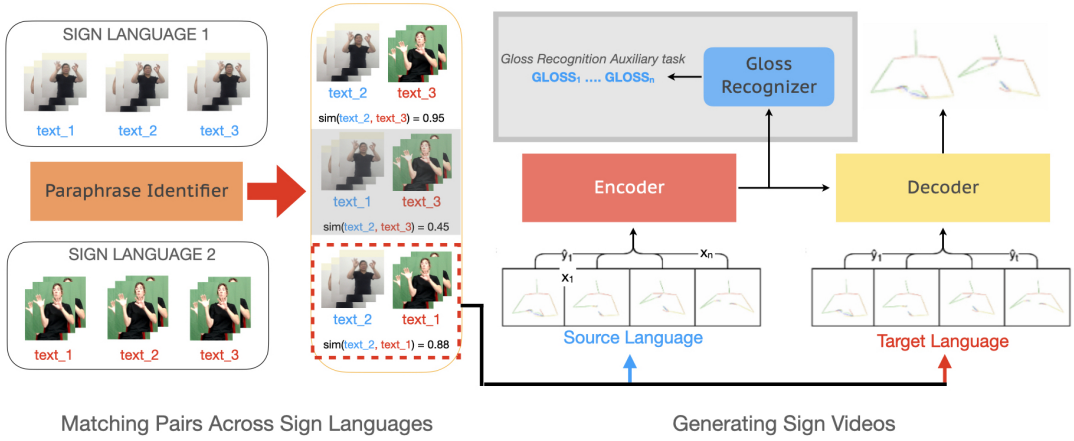

Figure 3: This figure shows the two stages of our approach: alignment and generation. We first have an alignment module where we identify paraphrases across languages and construct a parallel corpus using the textual modality for SLs (§4). In the next module, we generate signed videos using baseline models and prompt-tuned LLMs. We also introduce a multitasking model architecture with the additional gloss recognition auxiliary task (§5).

5 Our Generation Model

The Sign Language Translation (SLT) task generally involves multiple modular subtasks: sign-to-gloss, gloss-to-text, text-to-gloss, and gloss-to-sign translations. Our approach also uses this modular approach with additional constraints of the source and target SLs. In this section, we first introduce a transformer-based baseline model commonly used in the literature for sign-to-text translation (section §5.1). Then, to better utilize the multi-modality of the SLs, we further propose a multitasking model (Figure 3) that makes use of the corresponding glosses that come with the input sign sequence to introduce more in-domain knowledge into the encoder part. Lastly, we introduce an LLM-based translation model, which is posited to be pre-trained with extensive in-domain knowledge that can be better utilized for the gloss-to-gloss translation task.

5.1 Baseline Model

We present a baseline model based on the encoder-decoder architecture. The main goal of the cross-lingual sign language translation model is to transform a signing video from the source language into a 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 feed the encoded input skeletal joint sequences 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 $L_{MSE} = \frac{1}{T}\sum_{i=1}^{T}(y_i - \hat{y}_i)^2$.

5.2 Model with Gloss-based Multitasking

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 represents the sets of all viable paths for the glosses, 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.

5.3 LLMs for Sign Generation

As the third model family in our setup, we present a case study using an LLM for translation as proof of concept. Based on the previous model with an auxiliary gloss recognition task, this model also uses LLMs for the text-to-gloss and gloss-to-gloss translation tasks. It is postulated that large pretrained models have better semantic representation capabilities of text; hence, this is an effort to offload the gloss-based translations to LLMs. We describe the details of the text-to-gloss and gloss-to-gloss prompting setup in Appendix E.

6 Evaluations and Results

In this section, we present the automatic metrics we use and the evaluation paradigm for SLs. Then, we discuss the results of our experiments with these metrics.

6.1 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 6. Model implementation details and pipeline details are given in Appendices §C, and §D.

6.2 Automatic Results

We first train end-to-end baseline models on the six different language pairs. As shown in the first row of Table 5, 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 6). 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 models (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 $\rightarrow$ X tasks, we observe significant improvements across BLEU scores and ROUGE-L F1 scores. However, for CSL $\rightarrow$ X tasks, the gloss does not help much. One other difference is that for DGS $\rightarrow$ 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 when evaluating SLs with a larger vocabulary and less repetitive patterns of inputs, current back-translation metrics fail to evaluate the quality of the generated videos.

LLM results We present GPT-4 results for gloss-to-gloss generation, which was then evaluated against the ground truth of manually annotated glosses. The results from this experiment can be observed in Table 7. Here, it can be seen that the BLEU-3 and BLEU-4 scores are very low (around 0-5%). This shows that, at the sentence level, LLM translations may not be reliable replacements for rule-based gloss recognizers or transformer-based encoders. It can also be claimed that the BLEU metric cannot fully capture the semantic differences of glosses across SLs.

Further, it can be seen that ASL to DGS translation receives high scores, while ASL to CSL does not. Further, the opposite is true when CSL is translated to either of these languages. This may be due to the nature of the textual representations of signs using different glossing styles. As CSL glossing uses Chinese characters that are morphemes both in Chinese and CSL, while glossing for DGS and ASL uses letters which are phonetic units in English and German while a sign is a morpheme in DGS and ASL. This may be due to Chinese glosses packing more meaning and leading to better gloss-to-gloss matching and translations by the LLM. Overall, this case study on the LLMs as intermediate modules for translating glosses and texts for SLG shows that they are a new avenue to explore. Yet, we can not currently rely on their role and efficacy in sign-to-sign translation.

6.3 Qualitative Error Analysis

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 (Table 8).

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

Table 5: This table shows the results for our experiments on sign-to-sign translation (source → target) results on the test set, **bolded** lines are better results of the two models. Model performances vary across different languages pointing to the need for multilingual and aligned corpora.

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

Table 6: This is the table for back-translation results between within ground truth skeletal joints, glosses, and texts. These are upper bounds for the performance of the back-translation model itself.

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 7: Gloss translations using GPT-4 (source → target) results on the test set.

One concern in Table 5 is the low $BLEU_4$ score of 0.00 for the ASL/DGS to CSL translation task. 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 accurate enough and may introduce errors in the training stage. For instance, in the second example in Table 8, 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 produce *milk* as one of the foods ordered/eaten. This can be due to the domain of CSL, which covers daily-life entities.

Meanwhile, for ASL and CSL to DGS generations, we can refer to the back-translated results to examine the generation quality. As illustrated in the third row of Table 8, even though "good night" is over-generated, the back-translation result matches the paired DGS target sentence.

7 Discussions

We discuss essential challenges that demand future efforts in sign-to-sign generation. One such challenge is the difficulty of cross-linguistic and cross-cultural alignments. Similar to spoken languages, SLs can be affected by the physical and cultural factors of the user population. Thus, sign representations are not unified across SLs. There have been initiatives in the sign language processing community to use SignWriting as an intermediary unified textual representation (Jiang et al., 2023), but this is still in the process of wide adoption. In any case, as a community, we need more data with diverse and sociocultural relevance in addition to limited weather forecast- or Bible-based datasets. For further discussion on the sociolinguistic aspects of this problem, please refer to Appendix §F.

Another point of discussion is surrounding automatic evaluation metrics for signing. Current evaluations are restricted to the back-translation results of the generated sign videos, which lack

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: This table presents a qualitative analysis of model outputs. For non-English texts, we provide English translations (underlined) at 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 5. We find that, although sometimes the automatically aligned pairs do not covey the identical meaning, our model can produce reasonable results and cover salient tokens. The examples are selected from DGS-CSL, ASL-CSL, and ASL-DGS from top to bottom.

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 SL research community. Moreover, the fact that there are significantly few publicly available resources for SL with glosses limited our choice and scope of datasets to the PHOENIX-14T and CSL-Daily dataset. The ASL, such as How2sign (Duarte et al., 2021) came without oracle glosses, and we have to utilize imperfect SLT models to derive glosses from the original text, thus introducing more errors.

8 Conclusions and Future Work

In this work, we address the problem of cross-lingual SLT, introducing a challenge for automatic video translation between SLs with a focus on automatic, linguistically informed alignment. Our effort facilitates cross-lingual sign language understanding and offers insights into signed languages' social, cognitive, and linguistic nuances, improving our understanding of their use across communities.

We release the first automatically aligned corpus with cross-lingual pairs that span three SLs, which can serve as a benchmark for future research. We show that LLMs can also be instrumental as intermediary translators, yet further experimentation and incorporation are necessary to judge their efficacy. We demonstrate that incorporating the gloss information can assist in understanding the video, which highlights the need for using intermediary textual representation to integrate more structure or stronger signals for better translation systems. Future work could involve better body pose extraction techniques to better understand cross-linguistic and cross-cultural semantics and prosody of SLs. Also, text-to-video retrieval approaches (Duarte et al., 2022b; Zuo et al., 2023) can be used to verify video alignments across different SLs.

Ethics Statement

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 signers' privacy. Instead of using the original frames that could leak the personal information of signers, we extract human skeletal joints and generate videos accordingly. 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 machine learning-based NLP methods, our methods are likely to perform better for content that is better represented in training, leading to further bias against marginalized groups.

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 re-

sults demonstrate the need for better-constructed large-scale datasets with high-quality alignments and a focus on linguistics.

Another limitation of this work is automatic evaluation. The current back-translation technique is the only available automatic method in addition to human evaluation. In this work, we did not employ human evaluation due to limited available resources. Having human evaluation on this dataset is a possible future direction to validate the quality of the generations and alignments.

Acknowledgments

We would like to thank Vidya Ganesh for their contributions to LLM-related implementations. In addition, we would like to thank Kate Atwell for their feedback.

References

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

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, 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*.

Patrick Boudreault Christian Rathmann, Gaurav Mathur. 2000. Amsterdam manifesto.

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

Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. 2023. Machine translation from signed to spoken languages: state of the art and challenges. *Univ. Access Inf. Soc.*, pages 1–27.

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. *Preprint*, arXiv:2008.08143.

Biyi Fang, Jillian Co, and Mi Zhang. 2018. DeepASL: Enabling Ubiquitous and Non-Intrusive Word and Sentence-Level Sign Language Translation. *arXiv*.

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.

Shester Gueuwou, Sophie Siake, Colin Leong, and Mathias Müller. 2023. JWSign: A highly multilingual corpus of Bible translations for more diversity

in sign language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9907–9927, Singapore. Association for Computational Linguistics.

Anja Hiddinga and Onno Crasborn. 2011. Signed languages and globalization. *Language in Society*, 40(4):483–505.

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 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.

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*.

Zifan Jiang, Amit Moryossef, Mathias Müller, and Sarah Ebling. 2023. Machine translation between spoken languages and signed languages represented in SignWriting. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1706–1724, Dubrovnik, Croatia. 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.

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. 2023. Gloss-free end-to-end sign language translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12904–12916, Toronto, Canada. Association for Computational Linguistics.

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.

Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. Data augmentation for sign language gloss translation. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 1–11, Virtual. Association for Machine Translation in the Americas.

Mathias Müller, Malihe Alikhani, Eleftherios Avramidis, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Sarah Ebling, Cristina España-Bonet, Anne Göhring, Roman Grundkiewicz, Mert Inan, Zifan Jiang, Oscar Koller, Amit Moryossef, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, Katja Tissi, and Davy Van Landuyt. 2023a. Findings of the second WMT shared task on sign language translation (WMT-SLT23). In *Proceedings of the Eighth Conference on Machine Translation*, pages 68–94, Singapore. Association for Computational Linguistics.

Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023b. 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.

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.

Rachel Rosenstock. 2004. *An investigation of international sign: Analyzing structure and comprehension*. Ph.D. thesis, UMI.

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 Inan, Lorna Quandt, and Malihe Alikhani. 2023. Including facial expressions in contextual embeddings for sign language generation. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 1–10, Toronto, Canada. Association for Computational Linguistics.

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. *Preprint*, arXiv:2303.04671.

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.

Aoxiong Yin, Zhou Zhao, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. 2022. Mlslt: Towards multilingual sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5109–5119.

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 Challenges in Sign-to-Sign Translation

Although substantial efforts have been made in SLR and SLG, aligning different signed languages proves challenging due to their distinctiveness. The main challenges are discretizing continuous signing — i.e., delineating where a morphological unit of sign starts and ends — focusing on bilingual translations, and multilingual alignment. For one, there is a lack of research on accurately segmenting and recognizing discrete signs in continuous SL datasets with an open vocabulary. The current state-of-the-art SLR model (Renz et al., 2021) achieves a temporal boundary prediction F1 score of only 0.53 on the PHOENIX-14T dataset, limited to British Sign Language (BSL) and DGS. The largest word-level ASL dataset (Li et al., 2019) reports a Top-1 retrieval accuracy of 30% for a vocabulary of 2,000 words. Other SL corpora can have even smaller vocabularies, ranging from forty (DGS) (Ong et al., 2012) to a few hundred (CSL), with low recognition accuracies. These performances raise questions about the accuracy of isolated sign recognition for subsequent generation tasks with multiple languages.

Finally, in-the-wild alignment of multiple signed languages presents separate challenges from translation using a cleaned and parallel corpus. Multiple datasets have been recently made available for sign translation with multiple languages. However, most of these datasets present a subset of already aligned parallel videos in multiple signed languages. For instance, the Spreadthesign-Ten (SP-10) dataset introduced in Yin et al. (2022) presents parallel videos available for 10 SLs from the SpreadTheSign initiative. Even though a data collection initiative is key, curating already-aligned signed videos does not contain many of the challenges required to align these videos. The main challenge in alignment comes from curating sign videos that were not previously aligned. Addressing this challenge can benefit the data scarcity and parallelizing unaligned multilingual sign videos already existing in the literature.

B ASL gloss extraction

We retrained a sign language translation model that 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.

C 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 an 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 joint sequences. 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).

D Pipelining Details

We build the pipelines as follows: for each source sign language, we reuse the back-translation model that can recognize texts from the continuous skeletal joint sequences. For machine translation, we use Google Translate to translate 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 of 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 5 (BLEU-1 25.62 and ROUGE of 27.75).

E LLM Prompting Details

Text-to-Gloss The task setup of text-to-gloss translation changes from the previous models, which involve training and finetuning, to one that is done at the inference level using LLMs. We prompt GPT-4 with the default parameters (temperature= 1, max tokens= 256, top p= 1, frequency penalty= 0, and presence penalty= 0) to translate text to SL glosses. This task is text-only and assumes that LLMs already understand the textual representations of SLs. We measure the success of this text-to-gloss translation with LLMs by comparing the generated glosses with the ground truth using automatic metrics. In the case of How2Sign, which lacks glosses, we use the other model's outputs of glosses for a comparison.

Gloss-to-Gloss To translate across different SLs, we translate the intermediary textual representations using an LLM, which we call the gloss-to-gloss task. For gloss-to-gloss translation, we again prompt GPT-4 with the default parameters to translate between ASL, CSL, and DGS, exploring all possible translations of these glosses. The specific prompt we use is "`Translate the following American Sign Language glosses: [GLOSSES] to Chinese Sign Language glosses`". In addition, we incorporate sign language rules as additional context. Again, it is assumed that LLMs already contain *a priori* an understanding of the SL glosses. To test the performance of the gloss-to-gloss translations, we use automatic metrics when there are ground truth glosses both in the source and target SL. We use the best-performing other model's outputs when there are no ground truth glosses.

F Sociolinguistic Discussion

The absence of a global sociolinguistic understanding of signing, which fully accommodates the depth and complexities of regional SLs, poses significant barriers to cross-cultural and transnational communication, educational accessibility, and emergency communication (Hiddinga and Crasborn, 2011). Yet, this is a difficult issue to tackle and requires international collaborations and interdisciplinary initiatives. Further, there have been concerns, such as in the 'Amsterdam Manifesto,' regarding the lack of interpreters who can translate into multiple SLs at international conferences (Christian Rathmann, 2000; Rosenstock, 2004). In the manifesto, there is a call for abandoning the effort to interpret into multiple SLs, which faces logistical difficulties, and instead focus on interpreting into a widely used SL or International Sign (IS) language. Due to the interdisciplinary and social nature of these issues, a quick solution is not yet evident.