Multilingual Gloss-free Sign Language Translation: Towards Building a Sign Language Foundation Model

Sihan Tan[1,2], **Taro Miyazaki**[2], **Kazuhiro Nakadai**[1]
[1]Institute of Science Tokyo, [2]NHK Science & Technology Research Laboratories
{tansihan, nakadai} @ra.sc.e.titech.ac.jp
miyazaki.t-jw@nhk.or.jp

Abstract

Sign Language Translation (SLT) aims to convert sign language (SL) videos into spoken language text, thereby bridging the communication gap between the sign and the spoken community. While most existing works focus on translating a single sign language into a single spoken language (one-to-one SLT), leveraging multilingual resources could mitigate low-resource issues and enhance accessibility. However, multilingual SLT (MLSLT) remains unexplored due to language conflicts and alignment difficulties across SLs and spoken languages. To address these challenges, we propose a multilingual gloss-free model with dual CTC objectives for token-level SL identification and spoken text generation. Our model supports 10 SLs and handles one-to-one, many-to-one, and many-to-many SLT tasks, achieving competitive performance compared to state-of-the-art methods on three widely adopted benchmarks: multilingual SP-10, PHOENIX14T, and CSL-Daily.[1]

1 Introduction

Sign language translation (SLT) is a sophisticated cross-modal task that converts sign language (SL) into spoken language, serving as a crucial bridge between the deaf and hard-of-hearing community and the hearing world. Recent advancements in deep learning have significantly improved SLT performance, particularly through either *gloss-based* or *gloss-free*[2] approaches (Camgoz et al., 2018; Chen et al., 2022). While gloss-based methods benefit from intermediate linguistic supervision, they suffer from an information bottleneck, limiting their real-world applicability (Müller et al., 2023). In contrast, gloss-free methods directly learn from raw SL videos, making them more practical yet challenging. Despite progress in SLT, existing research predominantly focuses on translating a single SL into a single spoken language (*one-to-one* SLT). However, collecting large-scale annotated SL datasets is difficult. Leveraging multilingual resources could mitigate low-resource issues and enhance accessibility. Existing multilingual SLT (MLSLT) studies (Yin et al., 2022; Zhang et al., 2025) are mostly limited to specific datasets (*e.g.,* SP-10) or restricted to *many-to-one* translation. While MLSLT holds great potential, it often suffers from performance degradation due to **language conflicts**. For instance, we observed a BLEU drop of 1.50 in our universal training setting (In § 5 many-to-one). In addition, **alignment challenges** between SLs and spoken languages hinder the development of MLSLT. To address these limitations, we propose a multilingual gloss-free SLT model with token-level sign language identification (SLI), capable of handling diverse multilingual SLT scenarios. Our contributions are as follows:

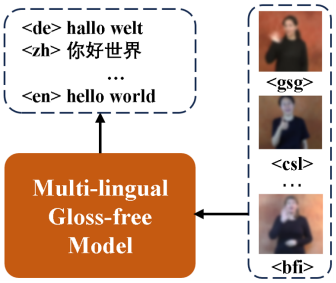

Figure 1: Overview of multilingual gloss-free model. Here, gsg = German Sign Language, csl = Chinese Sign Language, and bfi = British Sign Language.[3]

- We introduce *Sign2(LID+Text)*, a novel SLT approach that adopts dual CTC alignments: one with token-level SL IDs and the other with spoken languages, addressing the language conflicts and alignment challenges in MLSLT.

[1]Codes and model are available:https://github.com/Claire874/Gloss-free-MLSLT.

[2]Gloss is another written representation of sign language to help localize sign motions and simplify SLT tasks.

[3]We use the ISO639-3 and the ISO693-1 standard to represent sign language and spoken language.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 553–561
July 27 - August 1, 2025

- To the best of our knowledge, this is the first study to comprehensively explore one-to-one, many-to-one, and many-to-many gloss-free MLSLT, using multiple datasets (SP-10, PHOENIX14T, and CSL-Daily) and achieve state-of-the-art performance for each task.

2 Related work

Previous SLT studies mainly take a *cascading* or *end-to-end* approach. Cascading SLT, such as Sign2Gloss2Text, introduces gloss as intermediate supervision, simplifying SLT into two stages: sign language recognition (Sign2Gloss) and gloss-to-text translation (Gloss2Text) (Yin and Read, 2020). In contrast, end-to-end SLT directly converts sign videos into spoken texts. Camgoz et al. (2018) first proposed Sign2Text; however, it underperformed cascading SLT due to the challenging sign-text alignment. Later, Sign2Text was integrated with multi-task learning into Sign2(Gloss+Text) to alleviate the alignment issue (Camgoz et al., 2020). Recent work further advanced gloss-free SLT. Hamidullah et al. (2024) improved performance by introducing sentence embeddings as supervisions. In addition, large language models (LLMs) opened a new path for gloss-free SLT (Wong et al., 2024; Gong et al., 2024; Chen et al., 2024), but their applicability to multilingual settings is limited.

What is the current status of MLSLT? MLSLT remains underexplored due to several challenges. First, SLT itself involves the complex alignment between SL and spoken text, as SLs rely on fine-grained articulations such as finger spelling, palm orientation, and non-manual features (Liddell and Johnson, 1989). Second, language conflicts arise when training a unified model across diverse SLs. While certain SLs exhibit similarities, others differ greatly in structure and lexicon.[4] (Wei and Chen, 2023; Zhang et al., 2025). An intuitive solution is to introduce utterance-level SLI (Gebre et al., 2013); however, it is an ill-defined task, as models can learn to identify signers for particular SLs (Jiang et al., 2024). Inspired by token-level language identification in multilingual automatic speech recognition (ASR) (Chen et al., 2023), token-level SLI could offer a more flexible solution. Beyond resolving language conflicts, it could aid the model in mapping SLs to a large multilingual text space by providing fine-grained language cues throughout the sequence. Lastly, MLSLT suffers from a lack of large-scale datasets, and despite recent efforts (Yin et al., 2022; Gueuwou et al., 2023; Tanzer, 2024), data resources remain scarce.

[4]Examples are provided in Appendix A.

3 Method

To address the language conflict and alignment difficulties in MLSLT, we propose *Sign2(LID+Text)*, a novel approach that predicts token-level SL IDs (LID_{tok}) and translates SLs into spoken languages. Unlike previous studies (Gebre et al., 2013; Jiang et al., 2024) which predict an utterance-level LID (*e.g.,* a single label <ase> and <en> for American sign language), we introduce two auxiliary CTC objectives (Graves et al., 2006) to explicitly supervise LID_{tok} and target spoken text alignment, enabling hierarchical encoding under a joint CTC/Attention framework, as illustrated in Figure 2. This allows the early encoder layers to focus on token-level SLI, while the later layers reorder the latent sign representations for translation using a text-oriented CTC objective (TxtCTC). Table 1 summarizes the defined tasks and labels.

Tasks	Labels
MLSLT (many-to-many)	<en> hello world
MLSLT (many-to-one)	hello world
SLT (one-to-one)	hello world
Token-level SLI	<ase> <ase> <ase>

Table 1: Training label examples of Sign2 (LID+Text). In token-level SLI, all tokens are replaced with LIDs, while utterance-level SLI contains a single LID label.

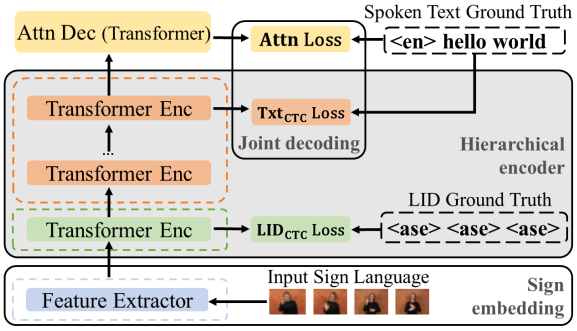

Figure 2: Overview of multilingual gloss-free model.

3.1 Feature extractor

In previous studies (Zhang et al., 2023; Tan et al., 2025), a pre-trained feature extractor on glosses has been used as sign embeddings, which is inherently designed for gloss-based SLT. Since our approach is gloss-free, we instead adopt a pre-trained feature extractor based on the SlowFastSign network (Ahn

et al., 2024) but trained on spoken texts. Pre-trained sign embeddings are further used to extract the sign feature $\mathcal{F}$ from the sign video sequence $\mathcal{V} = \{v_1, v_2, ..., v_{|\mathcal{V}|}\}$ consisting of $|\mathcal{V}|$ frames. This process is formulated as:

$$\mathcal{F} = \text{SignEmbedding}(\mathcal{V}), \quad (1)$$

where $\mathcal{F} = \{f_1, f_2, \ldots, f_{|\mathcal{F}|}\}$ denotes the extracted feature with $|\mathcal{F}|$ sign representations.

3.2 Hierarchical encoder

We employ a CTC-based hierarchical encoder, widely used in ASR (Sanabria and Metze, 2018) and machine translation (Yan et al., 2023), to facilitate multi-task learning and improve cross-modal alignments in MLSLT. The hierarchical encoder consists of two modules: an initial token-level SLI (Sign2LID) module and a subsequent Sign2Text module that reorders multilingual sign representations within a joint CTC/Attention framework, optimized with separate CTC objectives.

Sign2LID module is to predict the LID_{tok}, $\mathcal{I}_{\text{tok}} = \{i_1, i_2, ..., i_{|\mathcal{T}|}\}$ with the same length as the spoken text $\mathcal{T} = \{t_1, t_2, ..., t_{|\mathcal{T}|}\}$. We incorporate the LID_{tok}-oriented CTC loss as part of the multi-task Sign2(LID+Text) objective function.

$$\mathcal{L}_{\text{LID}} = -\log P_{\text{CTC}}(\mathcal{I}_{\text{tok}}|\mathcal{F}). \quad (2)$$

As in hierarchical conditioning, deeper layers handle increasingly complex predictions (Higuchi et al., 2022); the initial encoder layer suffices for the Sign2LID task. We assign this auxiliary task to the initial encoder layer, where the $|\mathcal{T}|$-length LID_{tok} sequence explicitly aligns the sign representations with each spoken word as

$$\mathbf{h}_{\text{int}} = \mathbf{Enc}_{\text{int}}(\mathcal{F}). \quad (3)$$

The intermediate sign representations $\mathbf{h}_{\text{int}}$ from the initial encoder layer, $\mathbf{Enc}_{\text{int}}$, are then forwarded to the subsequent encoder layers for Sign2Text.

Sign2Text module reorders the sign representations into the spoken text sequence. While CTC is typically constrained to monotonic alignments, neural network encoders allow for latent reordering (Zhang et al., 2022), enabling CTC to handle the non-monotonic alignment between the SL and spoken text. We apply TxtCTC to align the final encoder representation $\mathbf{h}_{\text{fin}}$ with the target spoken text sequence $\mathcal{T}$:

$$\mathcal{L}_{\text{Txt}} = -\log P_{\text{CTC}}(\mathcal{T}|\mathbf{h}_{\text{fin}}). \quad (4)$$

Following previous work (Yan et al., 2023), we frame our decoding process within a joint CTC/Attention setup, where the attention decoder plays a leading role in generating the output sequence, and the TxtCTC score provides auxiliary guidance during beam search. The overall training objective function jointly optimizes the hierarchical encoder and the attention decoder:

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{LID}} + \lambda_2\mathcal{L}_{\text{Txt}} + \lambda_3\mathcal{L}_{\text{Attn}}, \quad (5)$$

where $\mathcal{L}_{\text{Attn}}$ denotes the maximum likelihood estimation (MLE) loss for MLSLT, and λs control contributions of the Sign2LID, Sign2Text, and attention decoder objectives.

4 Experimental settings

To validate our proposed method, we conduct experiments on three tasks: one-to-one, many-to-one, and many-to-many SLT.

Datasets. We utilize three widely adopted datasets for our experiments: the multilingual SP-10 (Yin et al., 2022), RWTH-PHOENIX-2014T (PHOENIX14T) (Camgoz et al., 2018), and CSL-Daily (Zhou et al., 2021). SP-10 supports a broader range of tasks, featuring video recordings of 10 SLs from SpreadTheSign (Hilzensauer and Krammer, 2015). In contrast, PHOENIX14T and CSL-Daily are designed for one-to-one SLT. Appendix C provides statistics for the three datasets.

Implementation Details. We adopt a Transformer-based architecture within a joint CTC/Attention decoding framework (Tan et al., 2025). To evaluate the effectiveness of our method, we compare it with a vanilla Transformer baseline (Vaswani et al., 2017). Full implementation details and hyperparameters are provided in Appendix B.

Evaluation metrics. Following the previous studies, we evaluate performance using BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). BLEU is calculated through SacreBLEU (Post, 2018).

5 Results and Discussion

One-to-one evaluates the alignment capability of TxtCTC in standard single-pair SLT. Since Sign2LID is not utilized in this task, it is a purely gloss-free SLT setting. Tables 2 and 3 present results on SP-10, PHOENIX14T, and CSL-Daily, respectively. Our TxtCTC, integrated within the

joint CTC/Attention framework, achieves 1.71 and 2.24 BLEU improvements on PHOENIX14T and CSL-Daily. To further investigate the effect of TxtCTC, we present the token length distribution of PHOENIX14T and CSL-Daily along with the average BLEU4 score on each interval (see Figures 3 and 4). We observed that proposed TxtCTC within the joint CTC/Attention framework tends to be more effective for short and medium-length sentences. The impact of TxtCTC diminishes as sentence length increases. CTC tends to be more effective for segments with fewer ambiguities and clearer frame-to-token correspondences, which are more common in shorter sequences. For longer sequences, the model relies more on the attention mechanism to capture global context, which may naturally reduce the marginal contribution of the TxtCTC objective. In addition, Gloss-free SLT introduces an additional challenge that the input SL frame sequence is typically much longer than the spoken sentence. This inherent length difference increases the alignment difficulty, particularly for long sentences.

As no prior work reports one-to-one SLT results on SP-10 beyond English, we provide the first benchmark to facilitate future research.

Language Pairs	Dev		Test	
	BLEU	ROUGE	BLEU	ROUGE
csl → zh	8.79	35.59	7.32	32.40
ukl → uk	7.47	32.56	6.84	30.12
rsl → ru	6.20	31.79	4.23	28.98
icl → is	4.79	27.57	4.25	30.45
gsg → de	6.07	33.73	5.77	32.20
ise → it	5.88	30.13	4.76	27.91
bqn→bg	4.77	28.93	2.59	23.91
swl → sv	7.47	31.31	7.23	30.45
lls → lt	2.33	26.70	2.42	24.36
bfi → en	7.80	33.77	6.23	32.33

Table 2: One-to-one SLT results on the SP-10 dataset.

Many-to-one evaluates many-to-one SLT on the SP-10 dataset. Following previous studies (Yin et al., 2022; Zhang et al., 2025), we selected English as the target spoken language. A major challenge in the many-to-one setting is the language conflict, as confirmed by a preliminary experiment: the baseline many-to-one model suffers an average BLEU drop of 1.50 compared to individually trained one-to-one models (see Table 4 individual(10) and universal (1)). Instead, our *Sign2(LID+Text)* approach mitigates the language conflict and surpasses individual translation by 0.58 BLEU. Table 4 reports the results of each SL. Overall, our model outperforms previous ML-SLT systems. However, the limited target vocabulary (∼1.1k words) constrains further improvements. Data augmentation could be a promising way to address this limitation.

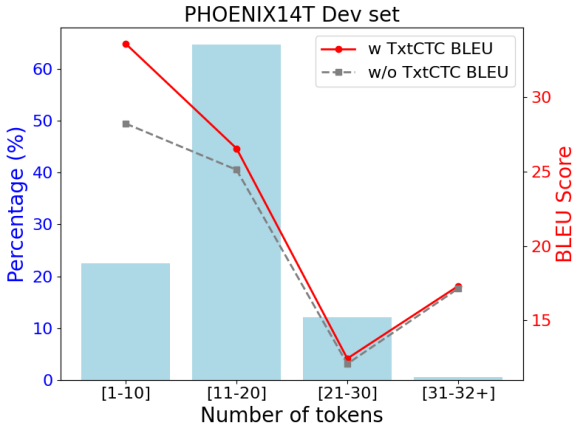

Figure 3: Average BLEU score on different token length intervals on PHOENIX14T.

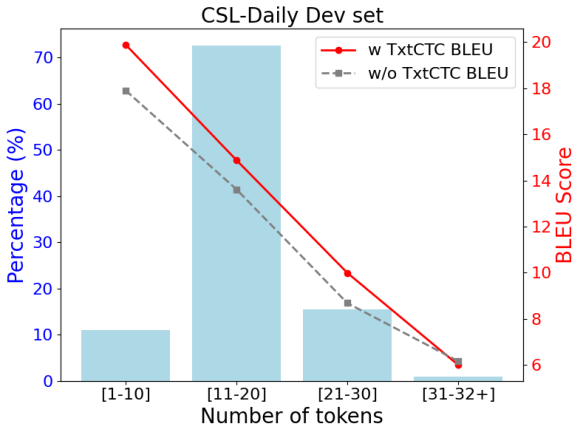

Figure 4: Average BLEU score on different token length intervals on CSL-Daily.

Many-to-many is the most challenging setting. We incrementally add language pairs based on their performance in Table 2, from highest to lowest on the dev set (see Appendix F). Table 5 presents the comparison with one-to-one SLT. Our many-to-many model maintains comparable performance as the number of language pairs increases to five. This stability benefits from cross-lingual information sharing, which reduces the reliance on large-scale data, particularly in low-resource SLT scenarios. To further validate LIDtok, we conduct an ablation study (see Appendix E) and find that LIDtok is especially effective under more challenging translation conditions. These results suggest that our many-to-many model provides a promising and scalable solution to mitigating the data scarcity

Methods	PHOENIX14T				CSL-Daily			
	Dev		Test		Dev		Test	
	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE
Gloss-free								
NSLT+Luong (Camgoz et al., 2018)	10.00	32.60	9.00	30.70	7.96	34.28	7.56	34.54
CSGCR (Zhao et al., 2022)	15.08	38.96	15.18	38.85	–	–	–	–
GFSLT-VLP (Zhou et al., 2023)	22.12	43.72	21.44	42.49	11.07	36.70	11.00	36.44
Sign2GPT (Wong et al., 2024)	–	–	22.52	48.90	–	–	15.40	42.36
Fla-LLM (Chen et al., 2024)	–	–	23.09	45.27	–	–	14.20	37.25
SignLLM (Gong et al., 2024)	**25.25**	47.23	23.40	44.49	12.23	39.18	**15.75**	39.91
Baseline	22.59	49.88	22.52	49.85	12.23	36.39	11.76	36.25
Ours w TxtCTC	24.18	**51.74**	**24.23**	**50.60**	**13.66**	**39.33**	14.18	**40.00**

Table 3: Experimental results on PHOENIX14T and CSL-Daily dataset for gloss-free SLT (one-to-one SLT).

Part/Metrics	Methods	csl	ukl	rsl	bqn	icl	gsg	ise	swl	lls	bfi	Mean
Dev / BLEU	Individual (10)	**8.02**	6.32	5.56	**4.88**	5.03	5.54	4.78	7.54	5.15	6.42	5.92
	Universal (1)	5.93	5.32	4.91	3.15	4.69	4.65	4.52	6.63	4.47	4.59	4.89
	Google Multi (Johnson et al., 2017)	2.46	3.14	2.93	2.21	3.44	2.71	3.18	2.89	1.81	3.49	2.83
	MLST (Yin et al., 2022)	5.16	5.42	4.95	3.28	**6.76**	5.18	**7.05**	6.33	**6.08**	7.03	5.72
	Ours	7.06	**6.77**	**7.38**	3.56	6.59	**5.59**	4.83	**7.76**	4.54	**7.79**	**6.19**
Dev / ROUGE	Individual (10)	36.60	33.61	30.05	27.70	31.51	32.47	29.88	35.80	31.44	32.57	28.53
	Universal (1)	32.74	31.64	31.37	26.19	30.18	29.26	30.12	33.19	31.16	30.22	30.61
	Google multi (Johnson et al., 2017)	28.50	28.93	30.01	24.66	29.91	29.75	28.33	31.01	27.7	32.42	29.12
	MLSLT (Yin et al., 2022)	34.59	34.04	31.62	**27.98**	35.29	33.50	**37.96**	36.02	**34.48**	37.25	34.27
	Ours	**36.77**	**34.86**	**36.02**	26.74	**35.64**	**34.95**	31.67	**37.97**	31.43	**37.56**	**34.36**
Test / BLEU	Individual (10)	**6.24**	4.00	3.69	**3.63**	3.30	3.77	3.40	6.21	4.49	6.23	4.60
	Universal (1)	2.72	2.36	2.19	3.02	4.14	3.31	1.19	3.94	3.20	4.70	3.10
	Google Multi (Johnson et al., 2017)	2.28	2.38	2.06	1.10	1.38	1.82	2.09	2.13	2.68	3.27	2.12
	MLSLT (Yin et al., 2022)	5.19	4.18	3.66	2.85	3.93	**4.97**	**6.70**	3.70	**5.72**	5.73	4.66
	Ours	5.92	**4.52**	**5.80**	2.93	**5.10**	4.65	5.00	**6.40**	5.13	**6.36**	**5.18**
Test / ROUGE	Individual (10)	34.51	30.93	30.90	**27.19**	28.00	30.19	27.66	33.92	31.36	35.73	31.04
	Universal (1)	29.57	29.03	28.96	26.66	31.75	30.12	26.73	31.11	30.33	32.41	29.67
	Google multi (Johnson et al., 2017)	29.37	28.63	29.57	23.95	28.53	29.36	29.30	29.83	30.03	30.76	28.93
	MLSLT (Yin et al., 2022)	33.33	**34.07**	31.54	25.75	33.25	32.13	**35.37**	33.09	**33.11**	35.34	32.70
	Ours	**35.18**	33.17	**34.17**	24.55	**34.57**	**32.83**	34.27	**34.98**	31.48	**36.16**	**33.14**

Table 4: Many-to-one SLT results on the SP-10 dataset, we select English as the target spoken language. The best performance is bolded, and the second-best is underlined.

problem and paves the way toward a unified SL foundation model in challenging SLT settings.

Language Pairs	One-to-one		Many-to-many	
	BLEU	ROUGE	BLEU	ROUGE
(2→2)	6.28	32.37	6.22	34.53
(3→3)	6.44	31.73	6.15	34.02
(4→4)	6.41	34.59	6.31	32.76
(5→5)	6.33	33.88	5.36	31.79
(6→6)	6.08	33.40	4.91	31.17
(7→7)	5.97	34.37	4.63	29.51
(8→8)	5.84	32.70	4.75	30.24
(9→9)	5.27	32.14	4.74	28.65
(10→10)	5.06	32.31	4.58	30.83

Table 5: One-to-one vs. many-to-many SLT.

6 Conclusion

To address language conflicts and alignment challenges in multilingual sign language translation (MLSLT), we proposed *Sign2(LID+Text)*, a multilingual gloss-free SLT model combining token-level sign language identification (Sign2LID) and sign-to-text CTC alignment (Sign2Text). Our approach achieved comparable performance with the state-of-the-art across one-to-one, many-to-one, and many-to-many SLT tasks on three widely adopted benchmarks, covering a total of 10 different sign languages (SLs). We showed that Sign2LID effectively mitigates language conflicts and Sign2Text improves sign-to-text alignment, especially for shorter and medium-length sequences. Our work encourages and lays the foundation for future exploration of large-scale multilingual gloss-free SLT and shows potential for enhancing cross-lingual SL processing, contributing to the development of a universal SL foundation model.

Limitations

The limitations of this work can be summarized as follows. First, data scarcity remains a major challenge. The SP-10 dataset is currently the only publicly available multilingual SLT corpus, and

as shown in Appendix C, each language in SP-10 contains only 830 training samples (8.3k overall), which is extremely small for training deep learning models. Moreover, as discussed in the many-to-one setting, the target language vocabulary is limited to approximately 1.1k words, further constraining the model's capacity to generate diverse outputs. Second, our many-to-many SLT evaluation is set to, at most, 10 language pairs. Extending the evaluation to the full 10×10 combinations poses a greater challenge and requires more computational resources. Future work will focus on scaling to more sign languages and more challenging settings.

Acknowledgment

We thank the (meta-)reviewers for their valuable feedback.

References

Junseok Ahn, Youngjoon Jang, and Joon Son Chung. 2024. Slowfast network for continuous sign language recognition. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3920–3924.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 7784–7793.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign Language Transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10023–10033.

William Chen, Brian Yan, Jiatong Shi, Yifan Peng, Soumi Maiti, and Shinji Watanabe. 2023. Improving massively multilingual asr with auxiliary ctc objectives. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie LIU, and Brian Mak. 2022. Two-stream network for sign language recognition and translation. In *Advances in Neural Information Processing Systems*, volume 35, pages 17043–17056. Curran Associates, Inc.

Zhigang Chen, Benjia Zhou, Jun Li, Jun Wan, Zhen Lei, Ning Jiang, Quan Lu, and Guoqing Zhao. 2024. Factorized learning assisted with large language model for gloss-free sign language translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7071–7081, Torino, Italia. ELRA and ICCL.

Binyam Gebrekidan Gebre, Peter Wittenburg, and Tom Heskes. 2013. Automatic sign language identification. In *2013 IEEE International Conference on Image Processing*, pages 2626–2630.

Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. 2024. Llms are good sign language translators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18362–18372.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *International Conference on Machine Learning*, page 369–376. Association for Computing Machinery.

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.

Yasser Hamidullah, Josef van Genabith, and Cristina España-Bonet. 2024. Sign language translation with sentence embedding supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 425–434, Bangkok, Thailand. Association for Computational Linguistics.

Yosuke Higuchi, Keita Karube, Tetsuji Ogawa, and Tetsunori Kobayashi. 2022. Hierarchical conditional end-to-end asr with ctc and multi-granular subword units. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7797–7801. IEEE.

Marlene Hilzensauer and Klaudia Krammer. 2015. A multilingual dictionary for sign languages:" spreadthesign". In *ICERI2015 Proceedings*, pages 7826–7834. IATED.

Zifan Jiang, Gerard Sant, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling. 2024. SignCLIP: Connecting text and sign language by contrastive learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9171–9193, Miami, Florida, USA. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Scott K. Liddell and Robert E. Johnson. 1989. American sign language: The phonological base. *Sign Language Studies*, (64):195–278.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023. Considerations for meaningful sign language machine translation based on glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 682–693, Toronto, Canada. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ramon Sanabria and Florian Metze. 2018. Hierarchical multitask learning with ctc. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 485–490.

Sihan Tan, Taro Miyazaki, Nabeela Khan, and Kazuhiro Nakadai. 2025. Improvement in sign language translation using text CTC alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3255–3266, Abu Dhabi, UAE. Association for Computational Linguistics.

Garrett Tanzer. 2024. Fleurs-asl: Including american sign language in massively multilingual multitask evaluation. *Preprint*, arXiv:2408.13585.

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

Fangyun Wei and Yutong Chen. 2023. Improving continuous sign language recognition with cross-lingual signs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23612–23621.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2024. Sign2GPT: Leveraging large language models for gloss-free sign language translation. In *The Twelfth International Conference on Learning Representations*.

Brian Yan, Siddharth Dalmia, Yosuke Higuchi, Graham Neubig, Florian Metze, Alan W Black, and Shinji Watanabe. 2023. CTC alignments improve autoregressive translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1623–1639, Dubrovnik, Croatia. Association for Computational Linguistics.

Aoxiong Yin, Zhou Zhao, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. 2022. Mlslt: Towards multilingual sign language translation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5099–5109.

Kayo Yin and Jesse Read. 2020. Better sign language translation with STMC-transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Biao Zhang, Barry Haddow, and Rico Sennrich. 2022. Revisiting end-to-end speech-to-text translation from scratch. In *International Conference on Machine Learning*, pages 26193–26205. PMLR.

Biao Zhang, Mathias Müller, and Rico Sennrich. 2023. SLTUNET: A simple unified model for sign language translation. In *The Eleventh International Conference on Learning Representations*.

Ruiquan Zhang, Cong Hu, Pei Yu, and Yidong Chen. 2025. Improving multilingual sign language translation with automatically clustered language family information. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3579–3588, Abu Dhabi, UAE. Association for Computational Linguistics.

Jian Zhao, Weizhen Qi, Wengang Zhou, Nan Duan, Ming Zhou, and Houqiang Li. 2022. Conditional sentence generation and cross-modal reranking for sign language translation. *IEEE Transactions on Multimedia*, 24:2662–2672.

Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. 2023. Gloss-free sign language translation: Improving from visual-language pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20871–20881.

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

A Examples of sign language conflict

As shown in Figure 5, many sign languages share similarities in expressing certain concepts. For example, when signing *rain*, signers often mimic the shape of raindrops falling, which is relatively universal. However, for *evening*, although the core concept involves representing the sun setting, variations in expression still exist across different sign languages.

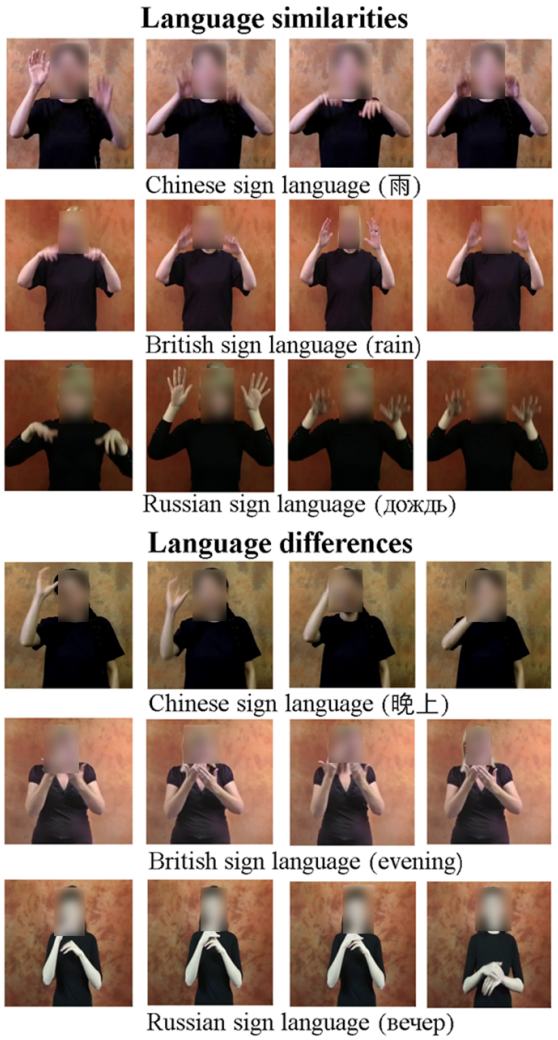

Figure 5: Sign language similarities and differences across languages. Sign videos are from SpreadTheSign. Note for privacy: we anonymize signers.

B Implement details

The Transformer model with a CTC/Attention setup uses a hidden size of 256 and a feed-forward dimension of 2048. Both the encoder and decoder have six layers. Training is conducted using the Xavier initializer and the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 1×10^{-3}. We train the model on an NVIDIA A100 (80GB) GPU with a batch size of 64. The total trainable #params is 39.48M. The training objective weights are set as $\lambda_1 = 1$, $\lambda_2 = 5$, and $\lambda_3 = 3$. When token-level LID is not used, λ_1 is set to 0.

C Summary of different SLT datasets

We summarize the statistics of the different datasets in Table 6, SP-10 consists of 10 different sign languages, with each having 830 training samples. In addition, Table 7 shows the involved sign languages with their abbreviations in SP-10.

Datasets	Lang	Statistics				
		#Signer	Vocab	#Train	#Dev	#Test
SP-10	multilingual	79	16.7k	8,300	1,420	2,021
PHOENIX14T	gsg	9	2.9k	7,096	519	642
CSL-Daily	csl	10	2.3k	18,401	1,077	1,176

Table 6: Statistics of SP-10, PHOENIX14T, and CSL-Daily datasets.

Languages	Abbr.
Chinese sign language	csl
Ukrainian sign language	ukl
Russian sign language	rsl
Icelandic sign language	icl
German sign language	gsg
Italian sign language	ise
Bulgarian sign language	bqn
Swedish sign language	swl
Lithuanian sign language	lls
British sign language	bfi

Table 7: Sign language abbreviations of the SP-10 dataset.

D The language conflict in SP-10

We performed a preliminary experiment to investigate language conflicts in multilingual SLT. The baseline model is adopted for individual (10) and universal (1) many-to-one SLT. The individual and universal translation performances are presented in Table 8. In general, the universal has an average 1.50 BLEU performance drop.

E Ablation study of token-level LID

Performance deteriorates as the number of languages involved in the many-to-many translation increases, and the translation task becomes more complex. As shown in Table 9, our method using token-level LID can suppress this deterioration and is effective in more complex translation settings.

Language Pairs	Individual (10) BLEU	Universal (1) BLEU	Variation
csl →en	6.24	2.72	−3.52
ukl → en	4.00	2.36	−1.64
rsl → en	3.69	2.19	−1.50
icl → en	3.30	4.14	+0.84
gsg → en	3.77	3.31	−0.46
ise → en	3.40	1.19	−2.21
bqn→en	3.63	3.02	−0.61
swl → en	6.21	3.94	−2.27
lls → en	4.49	3.20	−1.29
bfi → en	6.23	4.70	−1.53
Mean	**4.60**	**3.10**	**−1.50**

Table 8: Language conflicts in SP-10, we present the individual and universal translation results on the baseline.

Language Pairs	w/o LID_{tok} BLEU	w LID_{tok} BLEU
(2→2)	**7.48**	6.22
(3→3)	**6.50**	6.15
(4→4)	4.98	**6.31**
(5→5)	4.74	**5.36**
(6→6)	4.57	**4.91**
(7→7)	3.70	**4.63**
(8→8)	4.27	**4.75**
(9→9)	3.56	**4.74**
(10→10)	3.87	**4.58**

Table 9: Ablation study of token-level language ID (LID_{tok}) in many-to-many SLT.

F Details of language pairs in many-to-many SLT

Table 10 shows each language pair used in our many-to-many translation experiment.

Language Pairs	
(2→2)	(csl→zh) (bfi→en)
(3→3)	(csl→zh) (bfi→en) (swl→sv)
(4→4)	(csl→zh) (bfi→en) (swl→sv) (ukl→uk)
(5→5)	(csl→zh) (bfi→en) (swl→sv) (ukl→uk) (gsg→de)
(6→6)	(csl→zh) (bfi→en) (swl→sv) (ukl→uk) (gsg→de) (ise→it)
(7→7)	(csl→zh) (bfi→en) (swl→sv) (ukl→uk) (gsg→de) (ise→it) (rsl→ru)
(8→8)	(csl→zh) (bfi→en) (swl→sv) (ukl→uk) (gsg→de) (ise→it) (rsl→ru) (icl→is)
(9→9)	(csl→zh) (bfi→en) (swl→sv) (ukl→uk) (gsg→de) (ise→it) (rsl→ru) (icl→is) (bqn→bg)
(10→10)	(csl→zh) (bfi→en) (swl→sv) (ukl→uk) (gsg→de) (ise→it) (rsl→ru) (icl→is) (bqn→bg) (lls→lt)

Table 10: Language pairs used in many-to-many SLT