Rethinking Sign Language Translation: The Impact of Signer Dependence on Model Evaluation

Anonymous ACL submission

Abstract

Sign Language Translation has advanced with deep learning, yet evaluations remain signer-dependent, with overlapping signers in training, development, and test sets. This raises concerns about whether models truly generalise or rely on signer-specific features. To address this, signer-fold cross-validation is conducted on GFSLT-VLP, GASLT, and SignCL—three leading, publicly available, non-proprietary gloss-free sign language translation models, with SignCL being among the most prominent. Experiments are performed on two benchmarking datasets, CSL-Daily and PHOENIX14T. The results reveal a significant performance drop under signer-independent settings. On PHOENIX14T, GFSLT-VLP sees BLEU-4 fall from 21.44 to as low as 3.59 and ROUGE-L from 42.49 to 11.89; GASLT drops from a reported 15.74 to 8.26; and SignCL from 22.74 to 3.66. These findings highlight the substantial overestimation of SLT model performance when evaluations are conducted under signer-dependent assumptions. This work proposes two key recommendations: (1) adopting signer-independent evaluation protocols, and (2) restructuring datasets to include signer-independent splits.

1 Introduction

Sign Language Translation (SLT) is the task of automatically converting sign language videos into spoken or written language, enabling communication between the hearing-impaired and hearing communities. A major limitation in SLT is signer dependence in evaluation. Most SLT models are trained, validated and tested on dataset splits that do not enforce signer independence - indicating that the same signers appear across the training, development and test sets. This can lead to inflated performance metrics, as models may learn signer-specific patterns rather than generalising to unseen individuals. Signer independence refers to a model's ability to generalise to unseen signers, ensuring that performance is not biased toward individuals present in the training data (Liu et al., 2024; Mukushev et al., 2022; İnci Meliha Baytaş and İpek Erdoğan, 2024). Without explicitly accounting for signer variability, reported improvements in SLT models may reflect overfitting to signer-specific features.

The most widely used benchmark in SLT research, RWTH-PHOENIX-Weather-2014T (Phoenix14T) (Camgöz et al., 2018), represents this issue. Phoenix14T's default dataset split does not separate signers between training, development and test sets - making it inherently signer-dependent. Phoenix14T consists of weather forecast videos from German television channel PHOENIX, featuring nine professional sign language interpreters. It includes approximately 8,000 video sequences, spanning 11 hours of signing, along with their corresponding German translations and gloss annotations (Zhu et al., 2024). The dataset's gloss-level and sentence-level annotations make it valuable for evaluating both gloss-based and gloss-free SLT models.

Gloss-based models, which use manual gloss annotations as an intermediate representation, often achieve strong performance—but require costly manual annotations, as seen in the work of Yao et al. (2023), limiting their scalability. In response, recent research has increasingly explored gloss-free models (Zhou et al., 2023; Ye et al., 2024; Chen et al., 2024; Wong et al., 2024; Gong et al., 2024), which aim to map videos directly to spoken language, bypassing the need for gloss annotations—a resource that is often unavailable, particularly in low-resource settings. While these gloss-free approaches offer promising directions for broader applicability, their effectiveness remains constrained by signer-specific biases, as they are frequently evaluated on signer-dependent splits.

In addition to Phoenix14T, CSL-Daily (Zhou

et al., 2021) has emerged as a prominent benchmark dataset in recent SLT research. It provides over 20,000 high-resolution sign videos, annotated at both the gloss and sentence levels, and covers a wide range of daily-life topics such as travel, shopping, and medical care. The dataset features 10 native Chinese signers and supports both gloss-based and gloss-free SLT methods. Notably, CSL-Daily has been adopted by many recent SLT models (Zhou et al., 2023; Chen et al., 2024; Wong et al., 2024), establishing it as a standard evaluation benchmark alongside Phoenix14T. However, like Phoenix14T, its default dataset split does not enforce signer independence.

Despite the widespread use of Phoenix14T and CSL-Daily, to the authors' knowledge, no prior work has systematically tested the extent to which signer dependence affects the performance of a state-of-the-art SLT model. Furthermore, it is understood that no study to date has conducted signer-fold cross-validation across all signers present in the dataset. To mitigate against this issue, this research performs signer-fold cross-validation - ensuring that no signer appears in both the training, development and test sets. For this research, this study utilised the GFSLT-VLP (Zhou et al., 2023), GASLT (Yin et al., 2023), and SignCL (Ye et al., 2024) - the strongest publicly available gloss-free SLT model, to assess the impact of signer-independent training. The results reveal a significant drop in performance when evaluated under signer-independent conditions. The results of our experiments 3.2 demonstrates that signer overlap artificially inflates performance metrics, suggesting that reported improvements in SLT models may not accurately reflect real-world generalisation. Given that recent SLT models have been evaluated using the same dataset distribution, similar performance degradation is likely across the field.

This study highlights a critical gap in current SLT evaluation methodologies and calls for a shift towards signer-independent evaluation protocols. The following sections discuss related work, describes the experimental setup, presents key findings and proposes strategies to incorporate signer dependence in future SLT research.

2 Related Works

Signer independence is a critical challenge in Sign Language Recognition (SLR) and SLT, referring to a model's ability to generalise across different signers rather than overfitting to signer-specific characteristics such as hand shape, motion style and articulation speed. As a result, commonly used evaluation metrics may overestimate the true performance of the model.

The problem of signer independence has been actively studied in SLR, with early work dating back to 2013 (Ni et al., 2013). However, this issue has received little attention in SLT, where evaluations remain largely signer-dependent.

Several gloss-based approaches, including Gloss-to-Text (G2T), Sign-to-Gloss $\rightarrow$ Gloss-to-Text (S2G$\rightarrow$G2T), Sign2Gloss2Text (S2G2T), and Sign-to(Gloss-to-Text) (S2(G2T)) (Camgoz et al., 2018), and other models such as STMC-Transformer (Yin and Read, 2020), SimulSLT (Yin et al., 2021), and Hierarchical Spatio-Temporal Graph Neural Network (HST-GNN) (Kan et al., 2022), have been evaluated in this signer-dependent setup. While they show incremental improvements, their evaluations do not measure how well they generalise to unseen signers. Recent models, such as SLT with Iterative Prototype (IP-SLT) (Yao et al., 2023) and Conditional Variational Autoencoder for SLT (CV-SLT) (Rui Zhao, 2024), continue to follow the same evaluation approach.

Similarly, gloss-free approaches, including S2T (Camgoz et al., 2020), NSLT (Orbay and Akarun, 2020), Temporal Semantic Pyramid for SLT (TSP-Net) (LI et al., 2020), and Gloss Attention for Gloss-free Sign Language Translation (GASLT) (Yin et al., 2023), have also been evaluated on signer-overlapping dataset splits. More recent gloss-free models, such as GFSLT-VLP (Zhou et al., 2023), Sign2GPT (Wong et al., 2024), and SignLLM (Gong et al., 2024), leverage vision-language pretraining and LLMs to improve translation performance — yet their evaluations remain signer-dependent. Newer approaches, such as contrastive learning and factorised learning in Factorised Learning Assisted with Large Language Model for Gloss-free Sign Language Translation (FLa-LLM) (Chen et al., 2024) and SignCL (Ye et al., 2024), also lack signer-independent testing, making it unclear whether their improvements stem from advances in SLT or overfitting to specific signers.

Prior research in SLR has demonstrated that signer-dependent training inflates performance metrics (Podder et al., 2023). However, SLT research has yet to focus on this issue, as Phoenix14T and CSL-Daily's default split remains the stan-

dard evaluation protocol. To investigate this issue, signer-fold cross-validation was applied on GFSLT-VLP (Zhou et al., 2023), on SignCL (Ye et al., 2024), and on GASLT (Yin et al., 2023). While these are not the absolute best-performing SLT models on Phoenix14T, they serve as the highest-performing accessible benchmarks for assessing signer-independent training.

3 Experiments and Results

3.1 Methodology

The default distribution of Phoenix14T consists of 7,022 videos in the training set, 269 videos in the development set and 966 videos in the test set - with nine signers overlapping across these splits. While this setup facilitates training, it allows models to exploit signer-specific features such as hand shape, and signing style, rather than learning generalisable representations for unseen signers.

To address this issue, signer-fold cross-validation was applied to the Phoenix14T dataset. Unlike the default split, signer-fold cross-validation ensures that no signers are shared across training, development, or test sets. The dataset was divided into nine folds, with each fold containing videos from one signer exclusively used for testing, another for development, and the remaining signers for training. The size of the training set varied across folds, ranging from 5,100 to 7,893 videos, as shown in Table 1. This setup provides a robust framework for evaluating how well models generalise to unseen signers.

Similarly, the CSL-Daily dataset was reorganised to support signer-independent evaluation. Due to its large size — over 20,000 video samples—a 20% subset of videos was sampled per signer to make training and evaluation computationally feasible. Ten signer folds were created, with training set sizes ranging from 2,313 to 3,665 videos, development sets from 154 to 1,478 videos, and test sets from 154 to 395 videos, as shown in Table 4.

To assess model performance, BLEU-4 and ROUGE-L was employed. These metrics have been commonly used in SLT studies to evaluate translation quality, making them the standard for benchmarking SLT models. Therefore, this study adopted them to ensure comparability with existing research.

BLEU-4 (Bilingual Evaluation Understudy) measures the precision of n-grams between the generated and reference translations while applying a brevity penalty to prevent overly short outputs (Papineni et al., 2002). The BLEU score is computed as:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{4} w_n \log p_n\right) \quad (1)$$

where p_n represents the precision of n-grams up to length 4, w_n is the weight assigned to each n-gram, and BP is the brevity penalty defined as:

$$BP = \left\{ \begin{array}{ll} 1, & ifc > r \\ e^{(1-r/c)}, & ifc \leq r \end{array} \right\} \quad (2)$$

where c is the length of the generated translation and r is the length of the reference translation.

ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) evaluates translation quality based on the longest common subsequence (LCS) between the generated and reference sentences (Lin, 2004). The ROUGE-L score is computed as:

$$ROUGE - L = \frac{LCS(X, Y)}{\max(|X|, |Y|)} \quad (3)$$

where $LCS(X, Y)$ represents the longest common subsequence between the candidate translation X and the reference Y, and $|X|$ and $|Y|$ denote their respective lengths.

3.2 Results and Analysis

3.2.1 Results on Phoenix14T

The results in Table 1 show a consistent performance drop in GFSLT-VLP's performance on Phoenix14T under signer-independent conditions. While the default split achieves BLEU-4 of 21.44 and ROUGE-L of 42.49, performance drops significantly in the signer-independent setting, with BLEU-4 ranging from 3.59 to 17.30 and ROUGE-L from 11.80 to 34.02. This indicates that the model relies heavily on signer-specific cues, as its performance declines when tested on unseen signers.

Performance varies across different signer folds. The lowest scores appear in Fold 8 (BLEU-4: 3.59, ROUGE-L: 11.80), while Fold 6 records the highest (BLEU-4: 17.30, ROUGE-L: 34.02), suggesting that some signers introduce more significant challenges for the model, likely due to differences in signing style, articulation, or dataset imbalance. Additionally, folds with smaller test sets tend to exhibit more extreme score variations. Fold 6, with

only 47 test samples, achieves the highest BLEU-4 and ROUGE-L scores, while Fold 8, with 966 test samples, records the lowest, indicating that test set size impacts variability.

GASLT shows a similar trend to GFSLT-VLP, though with generally lower scores in both BLEU-4 and ROUGE-L. Under the default split, GASLT achieves 15.74 BLEU-4 and 26.39 ROUGE-L, but its performance in signer-independent folds ranges from 2.58 to 10.74 BLEU-4 and from 9.79 to 29.15 ROUGE-L. These results suggest that GASLT is also sensitive to signer variation, though potentially less so than GFSLT-VLP in some folds (e.g., Fold 3). Interestingly, GASLT outperforms GFSLT-VLP in Fold 3 (BLEU-4: 10.26 vs. 10.10), suggesting model differences in how signer characteristics are handled.

When aggregating results across the 9 signer folds, GFSLT-VLP achieves a mean BLEU-4 of 10.53 with a relatively high standard deviation of 4.02, and a mean ROUGE-L of 26.70 with a standard deviation of 6.59. This indicates greater variability in performance, possibly due to higher sensitivity to signer-specific features. In contrast, GASLT shows a lower mean BLEU-4 of 10.24 and mean ROUGE-L of 26.46, but with smaller standard deviations of 1.66 and 2.92 respectively, reflecting more consistent performance across folds.

To assess the statistical significance of the observed differences, paired t-tests and Wilcoxon signed-rank tests were conducted across the 9 signer folds. As shown in Table 2, the differences in BLEU-4 and ROUGE-L between GFSLT-VLP and GASLT are not statistically significant (*p* > 0.05). This suggests that, despite GFSLT-VLP achieving slightly higher average scores, the two models perform comparably under signer-independent evaluation.

In addition to model comparisons, one-sample t-tests and Wilcoxon signed-rank tests were used to evaluate whether the signer-independent scores were significantly lower than the standard signer-dependent baseline. Using the default split scores as reference values, both models exhibited statistically significant reductions in BLEU-4 and ROUGE-L under signer-independent conditions (*p* < 0.05), as shown in Table 3. These results confirm that signer-independent evaluation presents a substantially greater challenge for current SLT models.

3.2.2 Results on CSL-Daily

Table 4 presents signer-fold cross-validation results on the CSL-Daily dataset, where a 20% subset was sampled per signer to ensure computational feasibility while preserving the original signer distribution.

Overall, GASLT exhibits low BLEU-4 and ROUGE-L scores in this signer-independent setting, with BLEU-4 ranging from 0.00 (Folds 1 and 10) to 4.33 (Fold 4), and ROUGE-L ranging from 11.22 to 22.30. The average performance across all 10 folds is 1.75 BLEU-4 and 18.01 ROUGE-L, which is higher than its performance on the default signer-dependent split for BLEU-4 (0.82), but lower for ROUGE-L (20.28). These results indicate a reliance on signer-specific information and a significant generalisation gap when evaluated on unseen signers.

The standard deviations across folds are 1.26 for BLEU-4 and 3.00 for ROUGE-L, suggesting moderate variability in translation performance depending on the signer pair. The higher variance in ROUGE-L implies more fluctuation in sentence-level content coverage, whereas the lower variance in BLEU-4—despite very low absolute scores—suggests relatively consistent n-gram matching performance at this low baseline.

4 Conclusion

This study highlights the limitations of signer-dependent evaluation in SLT and underscores the necessity of adopting signer-independent benchmarking protocols. The experiments with signer-fold cross-validation on the Phoenix14T and CSl-Daily datasets demonstrate a significant drop in translation performance when models are evaluated under signer-independent conditions. Specifically, the BLEU-4 and ROUGE-L scores of GFSLT-VLP, one of the best performing gloss-free SLT model, were substantially lower in signer-independent splits compared to the default dataset distribution. These results indicate that prior evaluations may have overestimated model performance by inadvertently allowing models to exploit signer-specific cues such as hand shape, motion patterns, and signing style.

Given that many recent SLT models have been assessed using the same signer-dependent dataset splits, it is highly likely that other models would experience similar performance degradation under signer-independent conditions. This raises con-

Table 1: Signer-Fold Cross-Validation Results on PHOENIX14T for GFSLT-VLP and GASLT models. Metrics are BLEU-4 and ROUGE-L. Final row reports mean ± standard deviation across the 9 folds.

Fold	Dev Signer	Test Signer	Train Size	Dev Size	Test Size	GFSLT-VLP BLEU-4	GFSLT-VLP ROUGE-L	GASLT BLEU-4	GASLT ROUGE-L
1	Signer08	Signer01	5,100	966	2,191	6.65	21.80	7.94	24.46
2	Signer01	Signer02	5,971	2,191	95	8.49	20.61	8.89	22.75
3	Signer05	Signer03	5,641	1,933	683	10.02	26.70	10.19	27.10
4	Signer03	Signer04	6,367	683	1,207	13.70	32.30	11.02	29.21
5	Signer07	Signer05	5,458	866	1,933	11.90	29.70	9.79	27.45
6	Signer04	Signer06	7,003	1,207	47	17.30	34.02	12.65	31.33
7	Signer06	Signer07	7,344	47	866	9.19	26.30	8.26	25.40
8	Signer09	Signer08	7,022	269	966	3.59	11.80	10.07	27.55
9	Signer02	Signer09	7,893	95	269	13.95	32.07	13.38	30.87
Mean ± Std (9 folds)						**10.53 ± 4.02**	**26.70 ± 6.59**	**10.24 ± 1.66**	**26.46 ± 2.92**
Default Split			**7,096**	**519**	**642**	**21.44**	**42.49**	**15.74**	**39.86**

SignCL results have been excluded from the table due to incomplete coverage across folds. As of now, experiments have been completed on 8 out of 9 folds, and the remaining are still running on a compute cluster. Results will be included once all folds are complete.

Table 2: Paired statistical tests comparing GFSLT-VLP and GASLT scores across 9 signer-independent folds. None of the differences are statistically significant (*p* > 0.05).

Metric	Test	Stat.	*p*
BLEU-4	Paired t-test	0.28	0.79
	Wilcoxon	17.00	0.57
ROUGE-L	Paired t-test	-0.62	0.55
	Wilcoxon	20.00	0.82

None of the differences between GFSLT-VLP and GASLT are statistically significant across signer folds, indicating comparable performance under signer-independent evaluation.

Table 3: One-sample tests comparing signer-independent scores to default signer-dependent performance on PHOENIX14T. All results are significant (*p* < 0.05).

Model	Metric	Test	Stat.	*p*
GFSLT-VLP	BLEU-4	t-test	-7.85	<0.001
		Wilcoxon	0.00	0.004
	ROUGE-L	t-test	-6.91	<0.001
		Wilcoxon	0.00	0.004
GASLT	BLEU-4	t-test	-8.89	<0.001
		Wilcoxon	0.00	0.004
	ROUGE-L	t-test	-13.16	<0.001
		Wilcoxon	0.00	0.004

One-sample t-tests and Wilcoxon signed-rank tests confirmed that signer-independent performance is significantly lower than the default signer-dependent baseline across all metrics and models.

cerns about the generalisability of existing SLT models and calls for a shift in evaluation methodologies. Without rigorous signer-independent testing, improvements in BLEU-4 and ROUGE-L scores may not accurately reflect a model's ability to generalise across diverse signers and real-world signing variability.

Based on the findings of this study, the following recommendations for future SLT research is proposed:

- **Adopt signer-independent evaluation protocols:** Future studies should enforce strict separation of signers across training, development, and test sets to ensure a more reliable measure of generalisation.
- **Expand signer-diverse benchmark datasets:** Current datasets, including Phoenix14T and CSL-Daily, should be restructured or supplemented with signer-independent splits to better reflect real-world variability.
- **Explore signer-agnostic methods for signer-independent SLT:** Given that RGB video input captures signer-specific visual details such as appearance, and hand size, it may introduce biases that hinder generalisation. Skeleton-based representations, which encode only keypoints, offer a more signer-agnostic alternative. Future work should investigate whether skeleton-based models can enhance performance in signer-independent settings.

In implementing these measures, the field can move towards more reliable and generalisable models, ultimately improving sign language translation systems for real-world applications.

Table 4: Signer-Fold Cross-Validation Results on CSL-Daily (20% Subset) for the GASLT model. Metrics are BLEU-4 and ROUGE-L.

Fold	Dev Signer	Test Signer	Train Size	Dev Size	Test Size	GASLT BLEU-4	GASLT ROUGE-L
0	Signer01	Signer02	2,495	154	1,478	1.33	18.85
1	Signer02	Signer03	3,663	310	154	0.00	11.22
2	Signer03	Signer04	3,475	342	310	0.67	17.07
3	Signer04	Signer05	3,456	329	342	1.59	19.09
4	Signer05	Signer06	3,600	198	329	4.33	22.30
5	Signer06	Signer07	3,665	264	198	1.24	18.40
6	Signer07	Signer08	3,542	321	264	2.58	17.36
7	Signer08	Signer09	3,411	395	321	2.93	18.93
8	Signer09	Signer10	3,396	336	395	2.07	21.80
9	Signer10	Signer01	2,313	1,478	336	0.00	15.12
Mean ± Std (10 folds)						**1.75 ± 1.26**	**18.01 ± 3.00**
Default Split			**18,401**	**1,077**	**1,176**	**0.82**	**20.28**

Results for GFSLT-VLP and SignCL have been excluded due to incomplete fold coverage. GFSLT-VLP results are not yet available for 2 out of 10 folds (folds 7 and 9), and all SignCL results are still pending. Remaining experiments are currently running on a compute cluster and will be included in the final version.

Limitations

While this study provides a comprehensive evaluation of signer independence in SLT using several gloss-free models, including GFSLT-VLP, SignCL, and GASLT, some limitations remain. Specifically, the evaluation was constrained to models with publicly available implementations. As a result, other potentially stronger gloss-free approaches could not be included due to the lack of accessible code or pretrained models. Future work should encourage open-source availability of top-performing models to facilitate fair and reproducible signer-independent evaluations.

Secondly, the study does not incorporate alternative input representations, such as skeleton-based features, which may be more robust to signer variability, though may still retain signer-specific information. Future research should explore how different input modalities impact signer-independent performance and whether alternative representations can mitigate signer dependence.

Third, this study did not investigate gloss-to-text translation tasks, which may help disentangle the contribution of signer identity from linguistic content. Exploring signer-independent performance for gloss-based models remains a valuable direction for future research.

Despite these limitations, the findings of this work highlight the critical need for signer-independent evaluation protocols and dataset restructuring in SLT research. Addressing these challenges will help ensure that SLT models generalise beyond specific individuals and better reflect real-world applications.

References

Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT.

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.

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. *Preprint*, arXiv:2403.12556.

Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. 2024. Llms are good sign language translators. *Preprint*, arXiv:2404.00925.

Jichao Kan, Kun Hu, Markus Hagenbuchner, Ah Chung Tsoi, Mohammed Bennamoun, and Zhiyong Wang. 2022. Sign language translation with hierarchical spatio-temporal graph neural network. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3367–3376.

DONGXU LI, Chenchen Xu, Xin Yu, Kaihao Zhang, Benjamin Swift, Hanna Suominen, and Hongdong Li. 2020. Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. In *Advances in Neural Information Processing Systems*, volume 33, pages 12034–12045. Curran Associates, Inc.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tianyu Liu, Tangfei Tao, Yizhe Zhao, Min Li, and Jieli Zhu. 2024. A signer-independent sign language recognition method for the single-frequency dataset. *Neurocomputing*, 582:127479.

Medet Mukushev, Aidyn Ubingazhibov, Aigerim Kydyrbekova, Alfarabi Imashev, Vadim Kimmelman, and Anara Sandygulova. 2022. Fluentsigners-50: A signer independent benchmark dataset for sign language processing. *PLOS ONE*, 17(9):1–18.

Xunbo Ni, Gangyi Ding, Xunran Ni, Xunchao Ni, Qiankun Jing, JianDong Ma, Peng Li, and Tianyu Huang. 2013. Signer-independent sign language recognition based on manifold and discriminative training. In *Information Computing and Applications*, pages 263–272, Berlin, Heidelberg. Springer Berlin Heidelberg.

Alptekin Orbay and Lale Akarun. 2020. Neural sign language translation by learning tokenization. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 222–228.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Kanchon Kanti Podder, Maymouna Ezeddin, Muhammad E. H. Chowdhury, Md. Shaheenur Islam Sumon, Anas M. Tahir, Mohamed Arselene Ayari, Proma Dutta, Amith Khandakar, Zaid Bin Mahbub, and Muhammad Abdul Kadir. 2023. Signer-independent arabic sign language recognition system using deep learning model. *Sensors*, 23(16).

Biao Fu Cong Hu Jinsong Su Yidong Chen Rui Zhao, Liang Zhang. 2024. Conditional variational autoencoder for sign language translation with cross-modal alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

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*.

Huijie Yao, Wengang Zhou, Hao Feng, Hezhen Hu, Hao Zhou, and Houqiang Li. 2023. Sign language translation with iterative prototype. *Preprint*, arXiv:2308.12191.

Jinhui Ye, Xing Wang, Wenxiang Jiao, Junwei Liang, and Hui Xiong. 2024. Improving gloss-free sign language translation by reducing representation density.

Aoxiong Yin, Zhou Zhao, Jinglin Liu, Weike Jin, Meng Zhang, K6Xingshan Zeng, and Xiaofei He. 2021. Simulslt: End-to-end simultaneous sign language translation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4118–4127.

Aoxiong Yin, Tianyun Zhong, Li Tang, Weike Jin, Tao Jin, and Zhou Zhao. 2023. Gloss attention for gloss-free sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2551–2562.

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.

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.

Qidan Zhu, Jing Li, Fei Yuan, Jiaojiao Fan, and Quan Gan. 2024. A chinese continuous sign language dataset based on complex environments. *Preprint*, arXiv:2409.11960.

İnci Meliha Baytaş and İpek Erdoğan. 2024. Signer-independent sign language recognition with feature disentanglement. *Turkish Journal of Electrical Engineering and Computer Sciences*, 32(3).