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

Keren Artiaga[1], Sabyasachi Kamila[2], Haithem Afli[1], Conor Lynch[3], Mohammed Hasanuzzaman[1,4]

[1]ADAPT Centre, Munster Technological University, Cork, Ireland
[2]Manipal Institute of Technology Bengaluru, MAHE, Manipal, India
[3]Nimbus Research Centre, Munster Technological University, Cork, Ireland
[4]EEECS, Queen's University Belfast, UK

Correspondence: keren.artiaga@adaptcentre.ie

Abstract

Sign Language Translation has advanced with deep learning, yet evaluations remain largely signer-dependent, with overlapping signers across train/dev/test. This raises concerns about whether models truly generalise or rely on signer-specific regularities. We perform signer-fold cross-validation on GFSLT-VLP, GASLT, and SignCL, three leading, publicly available, gloss-free SLT models, on CSL-Daily and PHOENIX14T. Under signer-independent evaluation, performance drops sharply: in PHOENIX14T, GFSLT-VLP falls from BLEU-4 21.44 to 3.59 and ROUGE-L 42.49 to 11.89; GASLT from 15.74 to 8.26; and SignCL from 22.74 to 3.66. We also observe that in CSL-Daily, multiple signers perform many target sentences, so common splits can place identical sentences in both training and test, inflating absolute scores by rewarding recall of recurring sentences rather than genuine generalisation. These findings indicate that signer-dependent evaluation can substantially overestimate SLT capability. We recommend: (1) adopting signer-independent protocols to ensure generalisation to unseen signers; (2) restructuring datasets to include explicit signer-independent, sentence-disjoint splits for consistent benchmarking; and (3) reporting both signer-dependent and signer-independent results together with train–test sentence overlap to improve transparency and comparability.

1 Introduction

Sign Language Translation (SLT) aims to convert sign language videos into spoken or written language, enabling communication between deaf and hearing communities. A key challenge in SLT is signer dependence, where models are often trained and evaluated on splits where the same signers appear in both training and test sets. This setup risks overestimating model performance, as models may exploit signer-specific patterns rather than learning to generalise to unseen individuals (Liu et al., 2024; Mukushev et al., 2022; İnci Meliha Baytaş and İpek Erdoğan, 2024).

The most widely used SLT benchmark, RWTH-PHOENIX-Weather-2014T (Phoenix14T) (Camgöz et al., 2018) [1], exemplifies this problem. Its default split includes overlapping signers across train, development, and test sets. Although it provides high-quality gloss and translation annotations across 8,000 weather forecast videos featuring nine signers, evaluations on this split do not reflect signer-independent performance. A similar issue exists in CSL-Daily (Zhou et al., 2021) [2], a large-scale Chinese SLT dataset covering daily topics with 10 signers, which has also become a standard benchmark (Zhou et al., 2023; Chen et al., 2024; Wong et al., 2024).

Many recent SLT models—including both gloss-based (Yao et al., 2023) and gloss-free approaches (Zhou et al., 2023; Ye et al., 2024; Chen et al., 2024; Wong et al., 2024; Gong et al., 2024)—have reported impressive gains. However, these gains are typically measured on signer-overlapping splits, leaving open the question of how well such models generalise to unseen signers.

To address this, we perform signer-fold cross-validation on Phoenix14T and CSL-Daily, evaluating three state-of-the-art gloss-free SLT models: GFSLT-VLP (Zhou et al., 2023), GASLT (Yin et al., 2023), and SignCL (Ye et al., 2024). Each fold withholds a signer entirely from training and development. Our results (Section 3.2) show that, on average, translation performance drops under signer-independent evaluation compared to default-split baselines. This highlights how signer-dependent

[1]RWTH-PHOENIX-Weather-2014T is distributed by the HLT & Pattern Recognition Group, RWTH Aachen University, Germany. The dataset is publicly available for research use

[2]CSL-Daily is released by the Visual Sign Language Research Group, USTC, under a non-exclusive, non-transferable agreement. Access requires signing a license restricting use to non-commercial academic research, prohibiting redistribution or commercial use, and preserving subject anonymity.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 18379–18391
November 4-9, 2025

evaluations can obscure the true generalisation ability of SLT models.

This study provides the first systematic analysis of signer-independent SLT performance using signer-fold cross-validation. We advocate for future SLT research to adopt signer-independent protocols to ensure robust and fair evaluation. The remainder of this paper reviews related work, describes the experimental setup, presents results and qualitative analyses, and discusses implications for future SLT development.

2 Related Works

Signer independence remains a critical challenge in Sign Language Recognition (SLR) and Sign Language Translation (SLT), where models often overfit to signer-specific traits such as hand shape, articulation, or signing speed. While this issue has been well-documented in SLR since early work in 2013 (Ni et al., 2013), it has received comparatively little attention in SLT.

Most existing SLT research—whether gloss-based or gloss-free—continues to evaluate models using signer-overlapping splits. Gloss-based pipelines, including Gloss-to-Text (G2T), Sign-to-Gloss → Gloss-to-Text (S2G→G2T), and related variants (Camgoz et al., 2018; Yin and Read, 2020; Yin et al., 2021; Kan et al., 2022), have shown steady performance improvements, but do not assess generalisation to unseen signers. This limitation persists in recent models such as IP-SLT (Yao et al., 2023) and CV-SLT (Rui Zhao, 2024), which continue to rely on signer-dependent splits.

Gloss-free SLT approaches, which bypass intermediate gloss representations, have gained traction for their scalability in low-resource settings. Early models such as S2T (Camgoz et al., 2020), NSLT (Orbay and Akarun, 2020), and TSPNet (LI et al., 2020) laid the groundwork, but similarly evaluated only on signer-overlapping data. Recent vision-language pretrained models—GFSLT-VLP (Zhou et al., 2023), GASLT (Yin et al., 2023), SignCL (Ye et al., 2024), Sign2GPT (Wong et al., 2024), and SignLLM (Gong et al., 2024)—have reported strong results on Phoenix14T and CSL-Daily. However, these evaluations remain constrained to the default splits, which do not isolate signer effects.

Although signer-dependent training has been shown to inflate performance in SLR (Podder et al., 2023), SLT studies have yet to investigate this systematically. No prior work, to our knowledge, has performed signer-fold cross-validation in SLT to evaluate robustness across all signers. Our study addresses this gap by applying signer-fold evaluation to three strong, publicly available gloss-free models: GFSLT-VLP, GASLT, and SignCL. This allows us to quantify how signer overlap in standard protocols may obscure the true generalisation ability of SLT systems.

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. We report both fold-specific results for each test signer and aggregate results computed by averaging scores across all folds.

Similarly, the CSL-Daily dataset was reorganised to support signer-independent evaluation. Ten signer folds were created, with training set sizes ranging from 12,490 to 18,338 videos, development sets from 771 to 7,391 videos, and test sets from 771 to 1,978 videos, as shown in Table 8. For this dataset as well, we report both per-fold scores and mean aggregate scores across all ten folds.

For SignCL and GFSLT-VLP, training each signer-based fold on PHOENIX14T took approximately 6–7 days, while on CSL-Daily it took around 14 days, depending on the dataset size for each fold. The models were trained on an NVIDIA Quadro RTX 6000 GPU with 24 GB of dedicated VRAM. To ensure consistency and comparability with prior work, all hyperparameter values were

retained from the original implementations of the models. For GASLT, training across all folds on PHOENIX14T took 2–3 days using the same type of GPU (NVIDIA Quadro RTX 6000) with 24 GB of dedicated VRAM.

To assess model performance, corpus-level BLEU-4 and ROUGE-L were employed. All models were evaluated using their original implementations: SignCL and GFSLT-VLP report corpus-level BLEU and ROUGE scores using the `nlg-eval` (Sharma et al., 2017) package, while GASLT uses the `sacreBLEU` (Post, 2018) package for corpus-level BLEU and the `mscoco_rouge`[3] package for ROUGE-L. .

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:

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

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:

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

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:

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

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] https://github.com/tylin/coco-caption/tree/master/pycocoevalcap

3.2 Results and Analysis

3.2.1 Results on Phoenix14T

Table 1 shows a clear performance gap between the default and signer-independent settings. GFSLT-VLP, for example, drops from 21.44 BLEU-4 and 42.49 ROUGE-L (default) to an average of 10.53 BLEU-4 and 26.14 ROUGE-L across signer folds, a relative drop of over 50%. GASLT and SignCL exhibit similar trends, with GASLT falling from 15.74 to 10.24 BLEU-4 and SignCL from 22.74 to just 4.18.

To contextualise these results with respect to sentence overlap, Table 2 reports sentence-level leakage for PHOENIX14T. Dev/test *unique* sentences overlap only marginally with training (typically $\approx$1–3% per fold), and the fraction of test sentences performed by $\geq$3 signers is very small. Thus, broad repetition is rare and evaluation inflation from repeated targets is minimal.

Despite the overall degradation, per-signer performance reveals interesting patterns. GFSLT-VLP performs best in Fold 6 (BLEU-4: 17.30) but worst in Fold 8 (BLEU-4: 3.59), suggesting that signer variability—such as articulation style or test-set size—strongly impacts results. GASLT also fluctuates across folds but with smaller standard deviations, indicating more consistent signer-agnostic behaviour. In Fold 3, for instance, GASLT slightly outperforms GFSLT-VLP (10.19 vs. 10.02 BLEU-4), hinting at different sensitivities to signer traits. SignCL, while strongest in the default setting, shows the weakest generalisation, averaging just 4.18 BLEU-4 across folds.

These results underscore the importance of evaluating models under signer-independent protocols to reveal robustness gaps that may be hidden in standard splits.

3.2.2 Qualitative Analysis for Phoenix14T

Analysis on the representative fold To further illustrate model behavior under signer-independent conditions, we present a detailed set of qualitative comparisons in Table 3. This table focuses on multiple examples from Fold 3 (Signer03), a signer selected for its representative performance across models. Fold 3 is near the mean performance for GFSLT-VLP and GASLT, and ranks fourth for SignCL, with a squared error of 0.616—only slightly higher than its best-performing folds (Fold 4: 0.0055, Fold 8: 0.0508, Fold 9: 0.3762). As such, it offers a balanced setting for controlled eval-

Table 1: Signer-Fold Cross-Validation Results on PHOENIX14T for GFSLT-VLP, GASLT, and SignCL models

Fold	Dev Signer	Test Signer	Train Size	Dev Size	Test Size	GFSLT-VLP		GASLT		SignCL	
						BLEU-4	ROUGE-L	BLEU-4	ROUGE-L	BLEU-4	ROUGE-L
1	Signer08	Signer01	5,100	966	2,191	6.65	21.80	7.94	24.46	2.58	9.79
2	Signer01	Signer02	5,971	2,191	95	8.49	20.61	8.89	22.75	4.65	13.74
3	Signer05	Signer03	5,641	1,933	683	10.02	26.70	10.19	27.10	4.81	14.95
4	Signer03	Signer04	6,367	683	1,207	13.70	32.30	11.02	29.21	4.22	13.22
5	Signer07	Signer05	5,458	866	1,933	11.90	29.70	9.79	27.45	3.66	12.76
6	Signer04	Signer06	7,003	1,207	47	17.30	34.02	12.65	31.33	5.43	15.50
7	Signer06	Signer07	7,344	47	866	9.19	26.30	8.26	25.40	3.05	12.13
8	Signer09	Signer08	7,022	269	966	3.59	11.80	10.07	27.55	4.40	13.19
9	Signer02	Signer09	7,893	95	269	13.95	32.07	13.38	30.87	4.79	13.16
				Mean ± Std (9 folds)		**10.53 ± 4.17**	**26.14 ± 7.10**	**10.24 ± 1.86**	**27.35 ± 2.85**	**4.18 ± 0.92**	**13.16 ± 1.64**
Default Split			**7,096**	**519**	**642**	**21.44**	**42.49**	**15.74**	**39.86**	**22.74**	**49.04**

Table 2: Sentence leakage per signer-fold on PHOENIX14T.

Fold	Dev	Test	Dev overlap	Test overlap	Dev leak (sent)	Test leak (sent)	Dev+Test leak (sent)	≥3 signers
1	Signer08	Signer01	15	31	1.59%	1.47%	1.12%	0.86%
2	Signer01	Signer02	32	3	1.52%	3.26%	1.46%	3.26%
3	Signer05	Signer03	21	15	1.11%	2.23%	1.18%	1.49%
4	Signer03	Signer04	12	17	1.79%	1.42%	1.34%	1.00%
5	Signer07	Signer05	18	19	2.15%	1.01%	1.00%	0.74%
6	Signer04	Signer06	20	1	1.67%	2.22%	1.69%	0.00%
7	Signer06	Signer07	1	20	2.22%	2.39%	2.38%	1.56%
8	Signer09	Signer08	7	16	2.60%	1.70%	1.65%	1.48%
9	Signer02	Signer09	3	7	3.26%	2.60%	2.77%	1.86%

Rates are percentages of *unique* dev/test sentences that also appear in train. "Dev+Test leak (sent)" is computed over the union of dev and test unique sentences. "≥3 signers" is the fraction of test sentences performed by at least three signers

uation of model outputs.

The examples in Table 3 show clear differences in how closely each model reproduces the reference translations. Across examples, SignCL is the least faithful to the input, often producing content that is either unrelated or directly contradictory to the reference.

For instance, when the reference is a simple greeting—"Dear viewers, good evening"—both GFSLT-VLP and GASLT reproduce this exactly, while SignCL inserts a closing phrase about having a nice evening. This additional material was not present in the input. In another case, the reference describes an approaching low-pressure system bringing heavy snowfall and stormy winds. Instead of capturing this, GFSLT-VLP replaces the event with thunderstorms, GASLT shifts focus to dry and cold conditions, and SignCL includes generic snowfall and temperature values that were not mentioned in the reference.

There are also cases where SignCL introduces weather events that contradict the reference. For example, when the input states that the night will be mostly clear, SignCL claims it will rain in specific regions. In contrast, GFSLT-VLP and GASLT correctly describe clear skies, with some elaboration. Furthermore, GASLT gives a more moderate description that partially aligns with the reference.

In several predictions, SignCL outputs templated phrases such as "and now the weather forecast for tomorrow" even when the reference contains no such framing. These insertions suggest reliance on memorised genre patterns rather than the input itself. GASLT and GFSLT-VLP are less prone to such hallucinations, but still deviate by generalising the weather type or omitting nouns such as region or adverbs such as intensity.

Analysis on the best performing folds To gain further insight into the qualitative behaviour of the models, Table 4 presents a side-by-side comparison of translations produced by GFSLT-VLP and SignCL on selected examples from Fold 06, which corresponds to the best-performing fold for both models. Note that GASLT achieves its best performance on a different fold and is analysed separately.

The examples in the table highlight key differences in how GFSLT-VLP and SignCL handle con-

Table 3: Qualitative comparison of GFSLT-VLP, GASLT, and SignCL predictions on weather forecast references from Fold 3 (Signer03) for Phoenix14T

Reference Translation	Predictions
liebe zuschauer guten abend *(Dear viewers, good evening)*	**GFSLT-VLP: liebe zuschauer guten abend** *(Dear viewers, good evening)* **GASLT: liebe zuschauer guten abend** *(Dear viewers, good evening)* **SignCL: ihnen einen schönen abend und machen sie es gut** *(Have a nice evening and take care)*
morgen kommt dann kräftig bewegung ins wetter mit einem tief das uns kräftige schneefälle bringen wird und stürmischem wind *(Tomorrow, a low-pressure system will bring heavy snowfall and stormy winds)*	**GFSLT-VLP: morgen wechseln sich freundliche momente mit dichten wolken ab dabei fallen schauer am nachmittag sind auch einzelne gewitter dabei** *(Tomorrow, pleasant moments will alternate with dense clouds, with showers. There will also be isolated thunderstorms in the afternoon.)* **GASLT: morgen bleibt es bei uns noch trocken und auch in den nächsten tagen da wird es ein bisschen kälter und der mitte und im süden zum teil frostig** *(Tomorrow it will remain dry and in the next few days it will be a bit colder and partly frosty in the middle and in the south)* **SignCL: am mittwoch ist es windig später vor allem im süden neue schneefälle auch am donnerstag schneit es gebietsweise noch etwas minus elf bis minus ein grad** *(On Wednesday it will be windy, later especially in the south new snowfall. On Thursday there will still be some local snowfalls, from –11 to –1 degrees)*
es klart verbreitet auf minus fünfundzwanzig grad können im osten erreicht werden im westen zum teil auch bis minus einundzwanzig grad *(It will clear up widely, with lows of –25°C in the east and –21°C in the west)*	**GFSLT-VLP: am tag abgesehen von einigen schleierwolken viel sonne im süden ist es fast wolkenlos nur über die küsten ziehen auch mal dichtere wolken** *(During the day, apart from a few wispy clouds, there will be plenty of sun in the south. It will be almost cloudless, with only denser clouds occasionally moving over the coast.)* **GASLT: im westen sind es heute nacht minus fünf grad sonst werte um den gefrierpunkt oder minus vier bis minus** *(In the west it is minus five degrees tonight, otherwise temperatures around freezing or minus four to minus)* **SignCL: im norden und nordwesten fällt morgen vormittag gebietsweise noch regen in der mitte regnet es nur hier und da etwas im süden und osten gibt es neben ein paar schauern auch wolkenlücken** *(Rain still falls tomorrow morning in the north and northwest; only scattered rain in the center; in the south and east, some gaps in the clouds between showers)*
in der nacht ist es verbreitet klar oder nur gering bewölkt *(Overnight it will be mostly clear or only lightly cloudy)*	**GFSLT-VLP: in der nacht ist es meist klar oder nur leicht bewölkt besonders in den flussniederungen in der mitte und des südens wabern wieder ein paar nebelfelder** *(At night it will be mostly clear or only slightly cloudy, especially in the river valleys in the middle and south a few patches of fog will appear again)* **GASLT: in der nacht klart es verbreitet auf stellenweise bildet sich nebel oder hochnebelartige bewölkt** *(At night it will clear up widely, in places fog or high fog-like clouds will form)* **SignCL: in der nacht regnet es stellenweise etwas zwischen rheinland pfalz und brandenburg** *(Overnight it will rain in places between Rhineland-Palatinate and Brandenburg)*

tent preservation, topical focus, and hallucination. GFSLT-VLP typically remains aligned with the input, although it sometimes generalises temporal framing or introduces inferred details such as temperature values. In contrast, SignCL frequently outputs broadcast-style templates—such as "und nun die wettervorhersage…" or opening greetings like "guten abend liebe zuschauer"—which are fluent but semantically unrelated to the input. While both models occasionally hallucinate formulaic broadcast language, GFSLT-VLP tends to preserve more of the original meteorological content, whereas SignCL often substitutes it entirely with generic phrasing. These observations suggest that GFSLT-VLP maintains a stronger grounding in the source input, whereas SignCL appears more susceptible to overfitting on surface-level patterns.

On the other hand, Table 5 presents a qualitative comparison between GASLT predictions and the corresponding human reference translations from the best-performing fold (fold 09). The examples highlight a range of model behaviors, including semantic drift (top rows) and accurate reproduction of fixed expressions such as weather report introductions (bottom rows). These cases illustrate GASLT's tendency to occasionally hallucinate or generalise content (e.g., replacing atmospheric descriptions with generic closings), while also showing its reliability on highly formulaic phrases.

Overall, GASLT performs reliably on stereotyped and repetitive structures, such as introductory phrases, and demonstrates fluency with common weather-related terminology. However, it often diverges from the reference in subtle but meaningful ways—omitting key meteorological details, hallucinating plausible-sounding but incorrect content, or shifting the topical focus.

Analysis on the worst performing folds Meanwhile. Table 6 illustrates key failure modes of GFSLT-VLP on the weakest-performing fold (Fold 08) of Phoenix14T. The most prominent issue is template overgeneration, with the model frequently defaulting to fixed phrases such as "und nun die wettervorhersage für morgen mittwoch den sech-

Table 4: Qualitative comparison between GFSLT-VLP and SignCL predictions and reference translations on selected examples from the best-performing fold (fold 06) for Phoenix14T

Reference Translation	Predictions
im süden beginnt die nacht klar auch sonst lösen sich die meisten wolken auf *(In the south the night begins clear and most of the clouds will also dissolve)*	**GFSLT-VLP:** im süden dürfte es wahrscheinlich bis zum abend trocken bleiben *(In the south it is likely to remain dry until the evening)* **SignCL:** und nun die wettervorhersage für morgen sonntag den sechsundzwanzigsten november *(and now the weather forecast for tomorrow, Sunday the twenty-sixth of November)*
nur im nordwesten fallen später schauer *(only in the northwest will showers fall later)*	**GFSLT-VLP:** ihnen noch einen schönen abend und machen sie es gut *(Have a nice evening and take care)* **SignCL:** guten abend liebe zuschauer *(good evening dear viewers)*
am tag breiten sich die wolken mit einzelnen schauern von der nordsee über die nordwesthälfte aus *(During the day, clouds with occasional showers spread from the North Sea over the northwestern half)*	**GFSLT-VLP:** am tag neunzehn grad ganz im norden und zwölf grad in teilen thüringens und sachsens *(during the day nineteen degrees in the far north and twelve degrees in parts of Thuringia and Saxony)* **SignCL:** und nun die wettervorhersage für morgen donnerstag den sechsundzwanzigsten november *(and now the weather forecast for tomorrow, Thursday, November 26th)*
zwischen pfalz und sachsen bleibt es am längsten freundlich südöstlich davon regnet es zeitweise *(Between Palatinate and Saxony it stays friendly the longest, southeast of that it rains intermittently)*	**GFSLT-VLP:** im südosten bleibt es bis dreizehn grad weitgehend trocken *(In the southeast it will remain largely dry up to thirteen degrees)* **SignCL:** und nun die wettervorhersage für morgen donnerstag den sechsundzwanzigsten november *(and now the weather forecast for tomorrow, Thursday, November 26th)*

Table 5: Qualitative comparison between GASLT and reference translations on selected examples from the best-performing fold (fold 09). for Phoenix14T

Reference Translation	GASLT Prediction
aber so wird es nicht bleiben *(but it won't stay that way)*	**jetzt wünsche ich ihnen noch einen schönen abend** *(Now I wish you a nice evening)*
an den folgetagen bestimmen tiefdruckgebiete das geschehen *(In the following days, low pressure areas determine what happens)*	**am tag ist es zunächst noch freundlich im süden** *(During the day it is initially still pleasant in the south)*
und nun die wettervorhersage für morgen samstag den dreizehnten februar *(and now the weather forecast for tomorrow, Saturday the thirteenth of February)*	**und nun die wettervorhersage für morgen samstag den dreizehnten februar** *(and now the weather forecast for tomorrow, Saturday the thirteenth of February)*
und nun die wettervorhersage für morgen dienstag den siebten juli *(and now the weather forecast for tomorrow, Tuesday, July 7th)*	**und nun die wettervorhersage für morgen dienstag den siebten juli** *(and now the weather forecast for tomorrow, Tuesday, July 7th)*

sundzwanzigsten november" regardless of the input, indicating memorisation rather than grounded translation. Additionally, predictions often drift semantically from the reference, replacing specific weather events (e.g., stormy gusts) with unrelated or generic descriptions (e.g., cloud cover). The model also exhibits low output diversity, repeatedly generating fallback phrases such as "am tag mal sonne mal wolken . . . " across distinct inputs. These patterns suggest a reliance on high-frequency patterns at the expense of semantic accuracy.

Table 6: Qualitative comparison between GFSLT-VLP predictions and reference translations on selected examples from the weakest-performing fold (Fold 08) for Phoenix14T

Reference	GFSLT-VLP Prediction
am dienstag wird es dann verbreitet freundlich *(On Tuesday it will be generally friendly)*	**im norden und westen ist es heute nacht schon stark bewölkt** *(In the north and west it is already very cloudy tonight)*
liebe zuschauer guten abend *(Dear viewers, good evening)*	**guten abend liebe zuschauer** *(Good evening, dear viewers)*
bei gewittern sind stürmische böen möglich *(During thunderstorms, stormy gusts are possible)*	**im nordwesten ist es meist stark bewölkt** *(In the northwest it is mostly cloudy)*
und nun die wettervorhersage für morgen sonntag den siebenundzwanzigsten märz *(and now the weather forecast for tomorrow, Sunday the twenty-seventh of March)*	**und nun die wettervorhersage für morgen mittwoch den sechsundzwanzigsten november** *(And now the weather forecast for tomorrow, Wednesday November 26th)*
am wochenende sorgen dann noch mildere luft und viel wind gebietsweise für zweistellige plusgrade *(At the weekend, milder air and strong winds will cause double-digit plus temperatures in some areas)*	**und nun die wettervorhersage für morgen mittwoch den sechsundzwanzigsten november** *(And now the weather forecast for tomorrow, Wednesday November 26th)*

For GASLT and SignCL, we qualitatively analyse reference–prediction pairs from Fold 1, which yielded the poorest performance for both models. As shown in Table 7, SignCL displays a consistent failure mode in which nearly all predictions collapse to variations of a single phrase—*"und nun die wettervorhersage für morgen samstag den neunundzwanzigsten november"*—regardless of the input content. This fallback behavior suggests that the model fails to generalise and instead defaults to a memorised or overused template under signer-independent conditions. By contrast, GASLT exhibits more diverse but still problematic error patterns. A common trend is semantic drift, where the prediction diverges from the reference to describe unrelated weather scenarios. Another frequent issue is temporal mismatch, with outputs shifting events to different times of day, or introducing exaggerated weather details not present in the reference. While GASLT avoids collapsing to a single repeated phrase, its tendency to hallucinate or distort key details highlights difficulties in maintaining

temporal and semantic fidelity in low-performing folds.

Table 7: Qualitative comparison between GASLT and SignCL predictions and reference translations on selected examples from Fold 01 of Phoenix14T

Reference Translation	Predictions
guten abend liebe zuschauer *(good evening dear viewers)*	**GASLT:** das war es für heute schönen abend noch *(that's it for today, have a lovely evening)* **SignCL:** und nun die wettervorhersage für morgen samstag den neunundzwanzigsten november *(and now the weather forecast for tomorrow, Saturday the twenty-ninth of November)*
tiefer luftdruck bestimmt unser wetter *(low air pressure determines our weather)*	**GASLT:** das tief das von großbritannien nach südskandinavien zieht *(the low that moves from Great Britain to southern Scandinavia)* **SignCL:** und nun die wettervorhersage für morgen samstag den neunundzwanzigsten november *(and now the weather forecast for tomorrow, Saturday the twenty-ninth of November)*
und es ist ein bisschen kühler es ist erträglicher es sind fünfundzwanzig bis dreißig grad noch an nord und ostsee naja da haben luft und wasser dieselbe temperatur zwanzig bis zweiundzwanzig grad *(and it's a bit cooler, it's more bearable; it's still 25–30°C at the North and Baltic Sea; there the air and water have the same temperature, 20–22°C)*	**GASLT:** aber es ist sehr windig so in den nächsten tagen bis zum wochenende vor allen dingen im norden und an der nordsee da bleibt es mit minus vier bis minus vier grad im norden *(but it will be very windy in the next few days until the weekend, especially in the north and at the North Sea, where it will remain at around minus four degrees in the north)* **SignCL:** und nun die wettervorhersage für morgen samstag den neunundzwanzigsten november *(and now the weather forecast for tomorrow, Saturday the twenty-ninth of November)*
im nordosten zum teil noch regen *(in the northeast partly still rain)*	**GASLT: sonst ist es meist stark bewölkt** *(otherwise it is mostly cloudy)* **SignCL:** und nun die wettervorhersage für morgen samstag den neunundzwanzigsten november *(and now the weather forecast for tomorrow, Saturday the twenty-ninth of November)*

3.2.3 Results on CSL-Daily

Table 8 presents signer-fold cross-validation results on the CSL-Daily dataset. GASLT exhibits generally low performance in both the signer-dependent and signer-independent settings: in the latter, BLEU-4 ranges from 0.63 (Fold 2) to 7.48 (Fold 9) and ROUGE-L from 12.30 to 30.86, with an average across all 10 folds of 3.63 BLEU-4 and 21.98 ROUGE-L (BLEU-4 lower than the default signer-dependent split at 4.07, while ROUGE-L is higher than 20.35). For SignCL, on the four signer-independent folds currently evaluated (Folds 1–4), BLEU-4 ranges from 35.32 to 68.70 and ROUGE-L from 52.07 to 80.77, with a preliminary mean of 52.60 BLEU-4 and 67.80 ROUGE-L. Meanwhile, on the default signer-dependent split, SignCL records 22.62 BLEU-4 and 16.16 ROUGE-L, which are significantly lower than any of the results from the signer-independent folds

Overall, CSL-Daily is challenging for generalisation and appears sensitive to signer-specific familiarity. Because many target sentences are repeated at least three times across signers in CSL-Daily[4], signer-only splits can place identical sentences in train and test—an overlap known to inflate evaluation by favouring memorisation over genuine generalisation (Elangovan et al., 2021). This enables *template exploitation*, where models reproduce memorised target-sentence templates rather than learn signer-robust grounding. Table 9 quantifies this: in most folds (2–9), dev/test unique sentences overlap almost completely with training ($\sim 100\%$). Many test sentences are also performed by ≥ 3 signers, indicating extensive repetition. We therefore recommend enforcing sentence-level disjointness and dataset *deduplication*, which prior work on language-model datasets (Lee et al., 2022) shows reduces memorisation and train–test overlap, leading to more faithful evaluation.

3.2.4 Qualitative Analysis for CSL-Daily

Analysis on the representative fold To better understand model behavior under signer-independent conditions in CSL-Daily, Table 10 presents qualitative examples from Fold 08. Similar to our analysis of Phoenix14T in Table 4, we selected this fold for one of the qualitative analyses because its performance is closest to the average across all folds, making it representative of typical model outputs.

Qualitative inspection of Fold 08 (Table 10) reveals variability that is consistent with sentence leakage. The exact reproduction of "没有水和空气，任何生物都不能生存。" *(Without water and air, no living thing can survive.)* may reflect success on clips whose target sentence appears a lot in training with other signers, as this sentence is repeated by five different signers. By contrast, other predictions show severe semantic divergence from the references. These errors are suggestive of reliance on memorised sentence-level templates and strong decoder priors rather than signer-robust grounding; when a test clip does not closely match

[4]CSL-Daily is a studio-recorded corpus created for SLT, whereas PHOENIX14T is derived from real broadcast content; this difference helps explain why exact target-sentence repetition is common in CSL-Daily but comparatively rare in PHOENIX14T (see Table 2).

Table 8: Signer-Fold Cross-Validation Results on CSL-Daily for GASLT and SignCL. Metrics are BLEU-4 and ROUGE-L.

Fold	Dev Signer	Test Signer	Train Size	Dev Size	Test Size	GASLT		SignCL	
						BLEU-4	ROUGE-L	BLEU-4	ROUGE-L
1	Signer01	Signer02	12,490	771	7,391	3.20	23.10	35.32	52.07
2	Signer02	Signer03	18,330	1,551	771	0.63	12.30	48.51	64.80
3	Signer03	Signer04	17,390	1,711	1,551	1.06	17.74	57.88	73.56
4	Signer04	Signer05	17,293	1,648	1,711	5.66	25.17	68.70	80.77
5	Signer05	Signer06	18,011	993	1,648	6.26	29.60	–	–
6	Signer06	Signer07	18,338	1,321	993	4.48	20.85	–	–
7	Signer07	Signer08	17,723	1,608	1,321	3.26	20.84	–	–
8	Signer08	Signer09	17,066	1,697	1,608	3.17	20.99	–	–
9	Signer09	Signer10	16,991	1,680	1,978	7.48	30.86	–	–
10	Signer10	Signer01	15,181	7,391	1,680	1.06	18.34	–	–
				Mean ± Std (10 folds)		**3.63 ± 2.34**	**21.98 ± 5.55**	**–**	**–**
Default Split			**18,401**	**1,077**	**1,176**	**4.07**	**20.35**	**22.62**	**16.16**

Due to limited compute, GASLT is reported on all 10 folds, while SignCL is available for Folds 1–4 only. The completed SignCL folds show a consistent pattern—higher than GASLT on the same folds; hence, additional folds should refine but not change the qualitative conclusion.

Table 9: Sentence leakage per signer-fold on CSL-Daily.

Fold	Dev	Test	Dev overlap	Test overlap	Dev leak (sent)	Test leak (sent)	Dev+Test leak (sent)	≥3 signers
1	Signer01	Signer02	7,310	756	99.17%	98.18%	99.17%	98.05%
2	Signer02	Signer03	769	1,549	99.87%	99.94%	99.94%	49.42%
3	Signer03	Signer04	1,550	1,711	100.00%	100.00%	100.00%	99.53%
4	Signer04	Signer05	1,711	1,638	100.00%	99.88%	99.94%	98.78%
5	Signer05	Signer06	1,638	993	99.88%	100.00%	99.91%	99.60%
6	Signer06	Signer07	993	1,322	100.00%	100.00%	100.00%	99.17%
7	Signer07	Signer08	1,322	1,606	100.00%	99.94%	99.97%	99.50%
8	Signer08	Signer09	1,606	1,970	99.94%	100.00%	99.97%	82.74%
9	Signer09	Signer10	1,970	1,680	100.00%	100.00%	100.00%	72.32%
10	Signer10	Signer01	1,215	6,859	72.32%	93.05%	93.05%	77.09%

Rates are percentages of *unique* dev/test sentences that also appear in train. "Dev+Test leak (sent)" is computed over the union of dev and test unique sentences. "≥3 signers" is the fraction of test sentences performed by at least three signers

a learned template, the model can default to high-probability phrases, yielding redundancy and occasional syntactic breakage.

Analysis on the best performing fold Table 11 presents qualitative examples from Fold 9. Despite being the highest-scoring fold for GASLT, most translations still exhibit serious deviations from the reference, including hallucinations, omissions, and semantic drift.

As shown in Table 11, most examples continue to demonstrate the typical failure of GASLT under signer-independent evaluation. The first two rows reveal hallucinations and semantic substitutions—e.g., generating "drinks" instead of "restaurant" or producing fluent but off-topic sentences. The third row presents a more severe hallucination, where the model outputs a completely unrelated sentence, indicating a breakdown in content grounding.

However, the final example in the table shows a perfect match between the reference and prediction—specifically, "在 追求 理想 的 过程 中 感到 很 快乐 。" *(He feels very happy in the process of pursuing his ideal.)* is reproduced verbatim by GASLT. While such an agreement can reflect good generalisation on a short, frequent sentence, it is also consistent with the sentence repetition in CSL-Daily, which encourages template exploitation.

Analysis on the weakest-performing fold To understand GASLT's limitations in signer-independent settings, we analyse outputs from Fold 2—its lowest-scoring fold on CSL-Daily. Table 12 presents representative prediction–reference pairs. The selected examples reveal severe semantic drift,

Table 10: Qualitative comparison between GASLT and reference translations on selected examples from fold (fold 08) for CSL-Daily.

Reference	Prediction
逆境里正确认识自我和完善自我是一件十分重要的事。 *(It is very important to correctly understand and improve oneself in adversity.)*	利用风力发电,既可节约能源,又能保护环境。 *(Using wind power to generate electricity can save energy and protect the environment.)*
这笔钱是预留金，租完之后我们会退给你。 *(This money is a reserve and we will return it to you after the rental.)*	你要什么运动? *(What sport do you want?)*
没有水和空气，任何生物都不能生存。 *(Without water and air, no living thing can survive.)*	没有水和空气，任何生物都不能生存。 *(Without water and air, no living thing can survive.)*
老师今天带我们去活动，我们排队秩序井然。 *(The teacher took us to the activity today, and we lined up in an orderly manner.)*	现在我们可以通过改革开放后,我们减少网络流量。 *(Now we can reduce network traffic through reform and opening up.)*

Table 11: Qualitative comparison between GASLT and reference translations on selected examples from the best-performing fold (fold 09) for CSL-Daily.

Reference	Prediction
学校附近有一家好吃的饭店。 *(There is a delicious restaurant near the school.)*	学校附近有很多饮料很好吃。 *(There are many delicious drinks near the school.)*
电影院里不要玩手机。 *(Don't play with your cell phone in the cinema.)*	这部电影的电影很漂亮,感到很漂亮。 *(The film's cinematic quality is beautiful and feels very beautiful.)*
不要嫌弃学习较差的同学，要热心帮助他们。 *(Don't look down on students who are not good at studying, but be eager to help them.)*	餐巾纸上都要注意安全。 *(Be careful with napkins.)*
他在追求理想的过程中感到很快乐。 *(He feels very happy in the process of pursuing his ideal.)*	他在追求理想的过程中感到很快乐。 *(He feels very happy in the process of pursuing his ideal.)*

with most predictions bearing little relevance to the input meaning. For instance, descriptive or instructional sentences such as *"He blushed with shyness when the teacher criticised him in person"* are replaced with vague or interrogative utterances like *"What do you do when?"*. This pattern holds across the broader test set: many GASLT predictions default to short, generic questions or unrelated dialogue-like phrases (e.g., *"Did you see it?"*, *"Where is the subway station?"*). These errors suggest that the model struggles to ground its outputs in input semantics, likely due to signer variability. In contrast to folds where content alignment is preserved, Fold 2 highlights how signer shift can trigger fallback behaviors and semantic disconnection.

Table 12: Qualitative comparison between GASLT predictions and reference translations on selected examples from Fold 2 of CSL-Daily

Reference	GASLT Prediction
他被老师当面批评，害羞地脸红。 *(He blushed with shyness when the teacher criticised him in person.)*	你有什么时候做什么? *(What do you do when?)*
报名须填写报名登记表。 *(To register, you must fill out the registration form.)*	地铁站在哪里? *(Where is the subway station ?)*
不懂就问！ *(If you don't understand, just ask!)*	你怎么? *(What about you?)*
新款手机不质量好，外观也很漂亮。 *(The new mobile phone is not only of good quality but also looks beautiful.)*	你有什么时候做什么? *(What do you do when?)*

4 Conclusion

This study highlights limitations in current SLT evaluation, particularly the field's reliance on signer-dependent protocols. Across folds and datasets, evaluation on unseen signers often leads to lower scores, although the magnitude and pattern of effects differ between PHOENIX14T and CSL-Daily. To strengthen evaluation and address dataset-specific confounds such as sentence repetition, we recommend: adopting signer-independent evaluation as a default; enforcing sentence-level disjointness when constructing splits (e.g., split over unique sentences first and then map to clips); and exploring signer-agnostic representations and training strategies that reduce template exploitation (e.g., frequency-aware sampling/downweighting of repeated targets). By addressing both signer dependence and sentence-level duplication, the field can move toward more generalisable and practically deployable SLT systems.

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 they 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.

Acknowledgments

Funding: This work was conducted with the financial support of the Science Foundation Ireland ADAPT Centre for Digital Content Technology (Grant No. 13/RC/2106). The ADAPT Centre's grant is co-funded under the European Regional Development Fund.

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.

Aparna Elangovan, Jiayuan He, and Karin Verspoor. 2021. Memorization vs. generalization : Quantifying data leakage in NLP performance evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1325–1335, Online. Association for Computational Linguistics.

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.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

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).

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, Belgium, Brussels. Association for Computational Linguistics.

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*.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799.

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.

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).

Appendix A. Statistical Tests on Signer-Specific Folds on Phoenix14T

We report paired and one-sample statistical tests to support the analysis of signer-independent evaluation presented in the main paper. Table 13 shows paired t-tests and Wilcoxon signed-rank tests across 9 signer-specific test folds (cf. Table 1). Table 14 compares signer-independent scores to the default signer-dependent baseline using one-sample tests.

The results show that SignCL significantly outperforms both GFSLT-VLP and GASLT in the signer-independent setting across all metrics. This is supported by both parametric (t-test) and non-parametric (Wilcoxon signed-rank) tests, with all p-values below the 0.05 threshold. In contrast, no significant difference is observed between GFSLT-VLP and GASLT, suggesting comparable generalization capabilities between these two models. The consistently low p-values and large test statistics in comparisons involving SignCL indicate a robust and reliable performance advantage on unseen signers.

Table 13: Paired tests comparing signer-independent performance across models. Statistically significant results ($\mathbf{p} < 0.05$) are bolded.

Comparison	Metric	Test	Stat.	*p*
GFSLT-VLP vs GASLT	BLEU-4	t-test	0.28	0.7878
		Wilcoxon	17.00	0.5147
	ROUGE-L	t-test	-0.62	0.5527
		Wilcoxon	20.00	0.7671
GFSLT-VLP vs SignCL	BLEU-4	t-test	**5.01**	**0.001**
		Wilcoxon	**1.00**	**0.0109**
	ROUGE-L	t-test	**5.76**	**0.0004**
		Wilcoxon	**1.00**	**0.0109**
GASLT vs SignCL	BLEU-4	t-test	**14.04**	**0.0**
		Wilcoxon	**0.00**	**0.0077**
	ROUGE-L	t-test	**16.91**	**0.0**
		Wilcoxon	**0.00**	**0.0077**

Paired comparisons highlight statistically significant differences between SignCL and the other models in both BLEU-4 and ROUGE-L metrics.

Table 14: One-sample tests comparing signer-independent scores to default signer-dependent performance on PHOENIX14T. All results are statistically significant ($\mathbf{p} < 0.05$).

Model	Metric	Test	Stat.	*p*
GFSLT-VLP	BLEU-4	t-test	-7.85	<0.001
		Wilcoxon	0.00	0.0077
	ROUGE-L	t-test	-6.91	<0.001
		Wilcoxon	0.00	0.0077
GASLT	BLEU-4	t-test	-8.89	<0.001
		Wilcoxon	0.00	0.0077
	ROUGE-L	t-test	-13.16	<0.001
		Wilcoxon	0.00	0.0077
SignCL	BLEU-4	t-test	-60.81	<0.001
		Wilcoxon	0.00	0.0077
	ROUGE-L	t-test	-65.51	<0.001
		Wilcoxon	0.00	0.0077

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

Appendix B. Analysis on differences in sentence complexity across signers

Phoenix14T sentence complexity

Although the folds exhibit some variation in the proportions of short and long sentences, Table 15 shows no consistent pattern that correlates with translation performance. Length categories were computed globally using quartiles of sentence length across the dataset, ensuring consistent classification across folds. For example, Fold 2 has the highest proportion of extra short sentences (44.21%) but achieves relatively low BLEU and ROUGE-L scores for GFSLT-VLP (8.49 / 20.61), SignCL (8.89 / 22.75), and GASLT (4.65 / 13.74). In contrast, Fold 6, which has a more balanced sentence length distribution, achieves some of the highest scores across all three models (GFSLT-VLP: 17.30 / 34.02, SignCL: 12.65 / 31.33, GASLT: 5.43 / 15.50). These results suggest that sentence length distribution is not a key explanatory factor for performance variation. Instead, the differences are more likely attributable to signer-specific traits such as articulation clarity, signing speed, or use of non-manual markers.

Table 15: Sentence length distribution across folds for Phoenix14T.

Fold	Extra Short (%)	Short (%)	Medium (%)	Long (%)
1	34.55	21.36	20.72	23.37
2	44.21	21.05	18.95	15.79
3	32.09	22.88	22.05	22.98
4	32.89	23.45	22.70	20.96
5	33.78	20.64	22.30	23.28
6	29.79	21.28	29.79	19.15
7	30.37	18.48	22.98	28.18
8	32.09	22.88	22.05	22.98
9	37.17	21.93	23.42	17.47

CSL-Daily sentence complexity

Table 16 reveals variation in sentence-length distributions across signer folds in CSL-Daily. However, for *GASLT* there is no consistent trend linking these distributions to translation performance. For instance, Fold 4 (Signer05 as the test signer) has the highest proportion of extra–short sentences (40.29%) but does not yield the highest scores (BLEU-4: 5.66, ROUGE-L: 25.17). Conversely, Fold 5 (Signer06), which has the highest proportion of long sentences (34.54%), performs better (BLEU-4: 6.26, ROUGE-L: 29.60). Fold 2 (Signer03), with relatively balanced sentence lengths, performs poorly (BLEU-4: 0.63, ROUGE-L: 12.30). These examples suggest that

sentence-length distribution is not a primary explanatory factor for performance differences in signer-independent evaluation for GASLT. Because SignCL's fold-wise evaluation is still incomplete, we avoid drawing parallel conclusions for that model. Instead, signer-specific properties (e.g., articulation clarity, visual variability) and dataset-specific factors—particularly *sentence repetition across many signers*, which can create sentence-level video–text duplication under signer-only splits—are more likely to influence observed performance. Folds that happen to include more high-frequency sentences performed by multiple signers may exhibit elevated scores independent of their length profiles.

Table 16: Sentence length distribution across signer folds for CSL-Daily.

Fold	Extra Short (%)	Short (%)	Medium (%)	Long (%)
1	21.14	23.61	31.78	23.48
2	22.63	26.24	30.69	20.44
3	25.64	25.82	27.45	21.09
4	40.29	25.91	21.36	12.44
5	13.70	21.35	30.41	34.54
6	22.09	18.68	29.12	30.11
7	23.20	23.63	27.55	25.62
8	41.91	31.34	20.93	5.81
9	21.79	20.60	28.45	29.17
10	27.34	25.22	26.92	20.51