SiLVERScore: Why Aren't Semantically-Aware Embeddings Used for Sign Language Generation Evaluation?

Anonymous ACL submission

Abstract

Evaluating sign language generation has traditionally relied on back-translation, where generated signs are converted into text and assessed using text-based metrics. However, this approach presents significant challenges: (i) it leads to substantial information loss, failing to capture the multimodal nature of sign language—such as facial expressions, spatial structure, and prosody—and (ii) errors introduced during back-translation propagate through the evaluation pipeline.

In this work, we propose SiLVERScore, a novel semantically-aware embedding-based evaluation metric that assesses sign language generation in a joint embedding space. Our contributions include: (1) identifying limitations of existing metrics, (2) introducing SiLVERScore for semantically-aware evaluation, (3) demonstrating its robustness to semantic and prosodic variations, and (4) exploring generalization challenges across datasets. SiLVERScore offers a step toward more reliable evaluation of sign language generation systems[1].

1 Introduction

The ability to automatically evaluate sign language generation is critical for advancing accessibility and inclusion for the deaf and hard of hearing community. Accurate evaluation ensures that generated sign language content meets the needs of users. However, the development of impactful, fully automated systems is hindered by the lack of effective evaluation methods (Liu et al., 2023). Ensuring that model outputs are aligned with human expectations requires robust evaluation metrics specifically tailored to sign language's multimodal nature.

In this work, we introduce SiLVERScore (**Si**gn **L**anguage **V**ideo **E**mbedding **R**epresentation **Score**), a novel embedding-based metric for evaluating sign language generation. SiLVERScore directly compares generated and reference signs within a joint embedding space, capturing semantic and prosodic features.

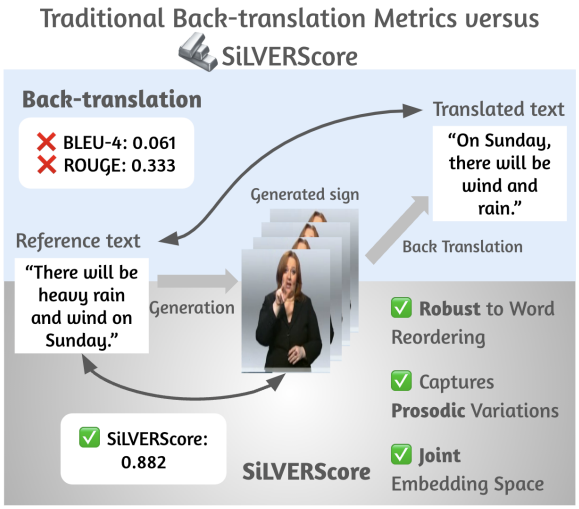

Figure 1: Comparison of evaluation methods for sign language generation. **Top**: Traditional back-translation-based metrics (e.g., BLEU, ROUGE) evaluate the generated sign by first translating it back to text, comparing the resulting text with the reference text. This approach ignores the actual sign and can lead to incorrect evaluations. **Bottom**: The proposed SiLVERScore uses embedding-based similarity to directly compare the generated sign with the reference text, ensuring a more accurate assessment of semantic alignment.

Automatically evaluating generated sign language remains challenging due to its unique multimodal linguistic nature, which incorporates facial expressions, manual markers, and spatiotemporal relationships into its prosody, iconicity, semantics, and pragmatics (Sandler, 2012; Liddell, 2003). Current evaluation methods rely on back-translation from visual to textual representations, which misaligns with the visual nature of sign language and leads to inaccuracies. While embedding-based metrics such as BLEURT (Sellam et al., 2020), BERTScore (Zhang* et al., 2020) and CLIPScore

[1]The GitHub link to the implementation and analysis will be disclosed after the review process to maintain anonymity.

(Hessel et al., 2021), have shown success in natural language processing, they have been underexplored for sign language evaluation. This limitation is primarily due to the scarcity and domain specificity of sign language datasets, which restrict the generalizability of sign embeddings. Our hypothesis is that these data limitations prevent the effective transfer of embedding-based metrics to sign language evaluation. To address this, we design SiLVERScore to be adaptable to individual datasets, ensuring robust evaluation despite limited data availability. This approach leads us to ask: *Can embedding-based metric offer a better alternative for evaluating sign language generation compared to back-translation?* Our work makes the following contributions:

1. We survey existing evaluation metrics for sign language generation and highlight their limitations (§ 2).
2. We introduce SiLVERScore, a novel semantically-aware embedding-based metric for evaluating sign language generation in a joint embedding space (§ 3).
3. We conduct prosodic and semantic tests to demonstrate that SiLVERScore outperforms traditional metrics, showing robustness to word reordering and prosodic variations (§ 4.2, § 4.3).
4. We perform a case study on generalization, the challenges of applying sign language models across different datasets and domains (§ 5).

2 Survey of Evaluation Metrics for Sign Language Processing

The evaluation of sign language generation systems has traditionally relied on back-translation approaches, first introduced by Camgoz et al. (2018). In these methods, a sign language translation model (typically trained by the authors) is used to convert the generated signs into text for evaluation. However, the absence of a standardized sign-to-text translation system complicates this approach, introducing unknown error propagation.

To address these issues, researchers have proposed several multimodal metrics. For instance, Dynamic Time Warping Mean Joint Error (Huang et al., 2021) aligns generated and ground truth poses to measure spatial-temporal accuracy and compute the mean joint error. While effective for motion similarity, it penalizes valid linguistic variations that differ in pose but maintain semantic meaning. Similarly, Fréchet Gesture Distance (Yoon et al., 2020), Fréchet Video Distance (Unterthiner et al., 2019), Fréchet Inception Distance (Heusel et al., 2017) compare gesture distributions but focus on physical similarity rather than semantics (Hwang et al., 2022; Xie et al., 2024; Hwang et al., 2024; Dong et al., 2024). In a visual-spatial SignWriting domain, signwriting-evaluation (Moryossef et al., 2024) was proposed as a metric designed for this by using its novel symbol distance metric using the Hungarian algorithm (Kuhn, 1955). A sign language translation metric, SignBLEU (Kim et al., 2024) aims to mitigate the significant information loss due to the simplification to a single sequence of text for evaluation. However, despite its improvements, both remain confined to the text-realm.

Embedding-based methods are promising due to their ability to capture multimodal elements and eliminate errors introduced by back-translation. Existing sign language embeddings, such as SignCLIP (Jiang et al., 2024), offer a foundation for embedding-based evaluation. However, they have not yet been widely adopted for evaluating sign language generation. This paper aims to bridge this gap by introducing and validating a semantically aware embedding-based evaluation metric tailored to sign language generation.

3 SiLVERScore

The objective of SiLVERScore is to evaluate generated sign language videos without requiring a reference video. This evaluation measures the alignment between a sign video and its corresponding text by comparing their similarity in a shared joint embedding space, trained to capture multimodal relationships. The similarities are computed using CiCo (Cheng et al., 2023), a model that leverages contrastive learning to align video and text representations. This approach addresses the alignment issues discussed in § 5 by using a sliding window mechanism to localize alignment between modalities.

We employ CiCo due to its framework that: (i) formulates sign language retrieval as a cross-lingual retrieval task; (ii) demonstrates state-of-the-art performance on benchmarks such as PHOENIX-2014T, CSL-Daily, and How2Sign; (iii) avoids reliance on pose estimation tools, eliminating dependency on pose extraction quality; and (iv) provides accessible code for implementation.

Model Details. The sign encoder processes sign videos using a sliding window mechanism to generate embeddings. This approach enables the model to handle continuous video streams without requiring explicit segmentation at test time. This encoder combines domain-agnostic features, captured by a pre-trained I3D network (Varol et al., 2021) on BSL-1K, with domain-aware features from the same network fine-tuned on PHOENIX-14T. The features are weighted and fused before being fed into a 12-layer Transformer initialized with CLIP's ViT-B encoder. The corresponding text is lowercased, byte pair encoded, and translated into English using Google Translate to align with the CLIP pretraining. The video and text embeddings are aligned through a contrastive learning objective with the InfoNCE loss. CiCo aligns video and text embeddings through a contrastive learning objective based on InfoNCE loss, which maximizes the similarity of matched video-text pairs while minimizing the similarity of unmatched pairs. This alignment is performed both globally across entire videos and texts and locally by retaining fine-grained mappings between video segments and individual text tokens. The resulting embeddings represent a semantically rich and temporally aware shared space that effectively captures the relationships between sign videos and their corresponding text annotations.

Global Similarity Calculation Global similarity is derived from a fine-grained similarity matrix $E \in R^{M \times L}$:

$$E(i,j) = S_i \cdot W_j^T, \tag{1}$$

where $S_i \in R^D$ and $W_j \in R^D$ represent video clip and word embeddings, respectively. To emphasize similarities, softmax re-weighting is applied:

$$E'(i,j) = \text{Softmax}(E(i,j)) \cdot E(i,j). \tag{2}$$

Row-wise summation followed by averaging yields the video-to-text similarity Z_{V2T}, while column-wise operations yield the text-to-video similarity Z_{T2V}.

In the implementation, the Z_{V2T} and Z_{T2V} similarities are equally weighted in the loss function. This equal weighting ensures that the global alignment of video-to-text and text-to-video pairs is equivalent, making it sufficient to use either Z_{V2T} or Z_{T2V} as the similarity metric. Without loss of generality, we use Z_{V2T} for our similarity metric.

Scaling for Interpretability To ensure comparability with metrics like BLEU and ROUGE, we follow a similar approach to CLIP-Score by scaling the embeddings with a weighting factor of 3, expanding the score distribution range to [0,100].

4 Experiments

To evaluate the effectiveness of SiLVERScore, we conduct multiple experiments to assess the performance compared to back-translation methods.

Dataset PHOENIX-14T dataset (Camgoz et al., 2018) is widely recognized as the benchmark dataset for sign language generation (Saunders et al., 2020, 2021; Viegas et al., 2023; Inan et al., 2022). It consists of German Sign Language weather forecast videos segmented into sentences, accompanied by corresponding German transcripts and sign-gloss annotations. The dataset includes 7,096 training samples, 519 validation samples, and 642 testing samples, recorded from 9 different signers.

Translation Model For the back translation model, we use the multi-stream keypoint attention network proposed by Guan et al., 2024, due to its state-of-the-art performance in sign language translation task of PHOENIX-14T dataset. This approach minimizes the error propagation caused by inaccuracies back translation.

Metrics We evaluate the quality of back-translated text using both rule-based and embedding-based metrics. For rule-based evaluation, we compute BLEU scores with sacreBLEU (Post, 2018) and ROUGE scores. For embedding-based evaluation, we use BLEURT (specifically BLEURT-20, Pu et al., 2021) and BERTScore (using the `bert-base-multilingual-cased` model to accommodate the German dataset; (Zhang* et al., 2020)). These metrics provide a benchmark for assessing the alignment quality of SiLVERScore in comparison to traditional back-translation evaluation methods.

4.1 Which metric can better distinguish between correct and incorrect video-text pairs?

4.1.1 Distribution of Metric Scores

To qualitatively evaluate the performance of different metrics, we analyze the kernel density plots in Figure 2. These plots illustrate the distribution of scores for correctly matched video-text pairs

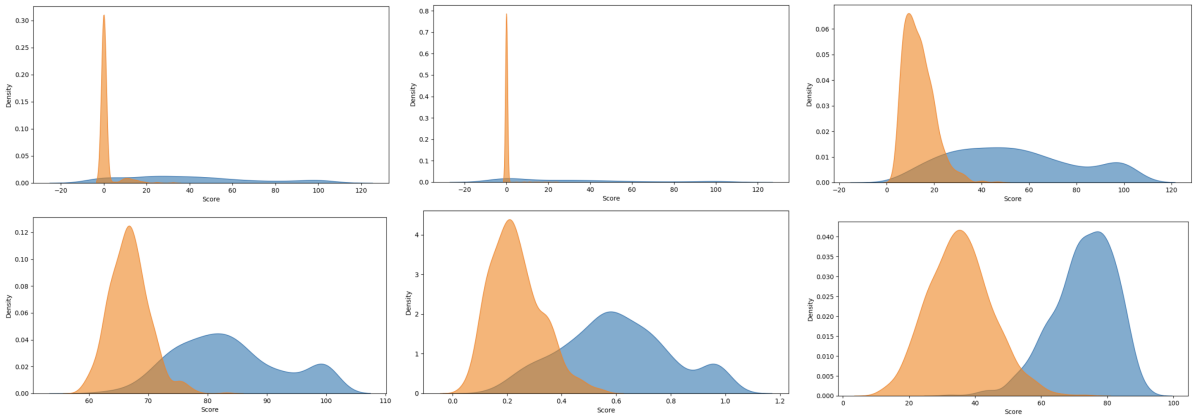

Figure 2: Kernel Density Plots for different metrics. **Top row (left to right, rule-based metrics):** BLEU-2, BLEU-3, ROUGE. **Bottom row (left to right, embedding-based metrics):** BERTScore, BLEURT, SiLVERScore. The blue curve represents the distribution of scores for matching indices (aligned pairs), while the orange curve represents different indices (misaligned pairs). SiLVERScore exhibits a clear separation between the two distributions, indicating a strong ability to distinguish aligned from misaligned pairs. In contrast, BLEU and ROUGE metrics show more overlap, reflecting their sensitivity to surface-level variations.

(blue curve) and randomly paired samples (orange curve). SiLVERScore shows a clear separation between the two distributions, with minimal overlap. This indicates its strong ability to distinguish aligned pairs from misaligned ones. In contrast, BLEU-2 exhibit significant overlap, particularly for lower score ranges, suggesting reduced discriminative power for this task. Similarly, the ROUGE plot shows partial separation but retains overlap between the two distributions. The BERTScore and BLEURT plots show improved separation compared to BLEU and ROUGE but still exhibit some overlap. The sharp distinction and density clustering of scores in the SiLVERScore plot indicate its effectiveness in capturing semantic alignment between video and text representations. The rest of the plots are in the Appendix A.

4.1.2 Quantifying overlap and separability

To complement the qualitative insights from the kernel density plots, we quantify the ability of each metric to distinguish between correctly aligned and randomly paired samples using overlap percentage and ROC AUC (Receiver Operating Characteristic Area Under the Curve). The results are summarized in Table 1.

Overlap percentage Overlap percentage measures how much the distributions of scores for correct and random pairs intersect. Lower overlap percentages indicate better discriminative power. Lower overlap percentages indicate better discriminative power.

Metric	Overlap (%)	ROC AUC
BLEU-1	53.74	0.95
BLEU-2	26.48	0.90
BLEU-3	38.94	0.81
BLEU-4	55.45	0.72
ROUGE	49.84	0.95
BERTScore	47.82	0.97
BLEURT	65.11	0.95
SiLVERScore	34.89	0.99

Table 1: Comparison of Overlap Percentages and ROC AUC for Various Metrics. SiLVERScore achieves the best overall performance with a low overlap of 34.89% and a high ROC AUC of 0.99.

Since each metric operates on a different scale, we applied Min-Max normalization to scale all metrics to the [0,1] range for a fair comparison.

From Table 1, BLEU-2 achieves the lowest overlap percentage (26.4798%). However, as observed in the kernel density plots, this low overlap does not translate to effective separability due to the dispersed and overlapping nature of the BLEU-2 distributions. SiLVERScore, with an overlap percentage of 34.8910%, shows clear separation in the density plots. The distributions are narrow and well-clustered, making the overlap region small and localized.

ROC AUC ROC AUC measures the metric's ability to distinguish between the two distributions. Higher ROC AUC values indicate better separability, with a maximum value of 1.0. SiLVERScore

achieves the highest ROC AUC of 0.9934, suggesting its superior performance in distinguishing aligned pairs. Despite BLEU-2 having a low overlap percentage, its ROC AUC is lower (0.9017) than SiLVERScore, confirming that its distributions are not well-separated. Overall, the results show that learned embedding-based metrics (SiLVERScore, BERTScore, BLEURT) outperform rule-based metrics in distinguishing between correctly aligned and misaligned video-text pairs.

4.2 Which metric captures semantic distinctions through targeted changes in the input?

4.2.1 Reordering

Rule-based metrics (BLEU and ROUGE) are inherently sensitive to the exact ordering of words, even when the overall meaning remains unchanged. To demonstrate this sensitivity, we designed an experiment where GPT-4o was used to reorder the words in sentences while preserving their meaning. The exact prompt provided to GPT-4o was:

> *Reorder the words in the following sentence while keeping the meaning the same: {text} Reordered sentence:*

Kernel density plot The kernel density plot (Figure 3) illustrates how different metrics respond to surface-level changes, specifically word reordering, while preserving the semantic meaning. SiLVERScore exhibits the highest score distribution, suggesting its robustness to reordering and its ability to capture semantic content. In contrast, BLEU and ROUGE display sharp peaks and narrower distributions concentrated in the lower score range. This pattern exhibits a clear distinction between rule-based and embedding-based metrics.

Quantifying overlap and separability In this experiment, the scores are computed by comparing the ground-truth references with their corresponding hypotheses. While these hypotheses may contain errors, they represent the best available approximations of the ground truth. By computing the ROC AUC between reordered pairs and reference pairs, we measure each metric's ability to distinguish between semantically similar and dissimilar pairs. Lower ROC AUC values indicate that the metric maintains its scores despite reordering, reflecting robustness to surface-level variations.

From Table 2, we observe that BLEU and ROUGE show significant drops in overlap percentages and higher ROC AUC values, indicating their sensitivity to word order. In contrast, SiLVERScore achieves the highest overlap (83.49%) and a relatively low ROC AUC (0.60), suggesting it better maintains robustness to reordering.

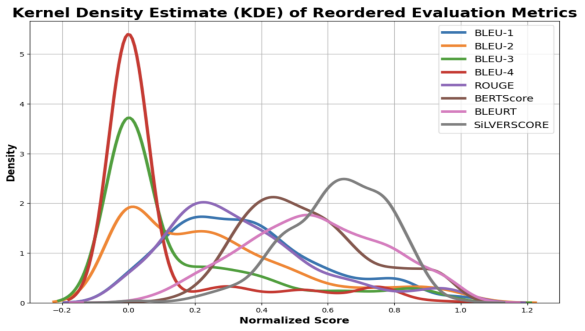

Figure 3: Kernel Density Estimate (KDE) plot comparing the score distributions of different evaluation metrics when applied reordered hypotheses. SiLVERScore, BERTScore, and BLEURT show broader distributions and higher overlap, while rule-based metrics such as BLEU and ROUGE exhibit sharp peaks at lower scores. This indicates their sensitivity to surface-level word order changes.

It is important to note that the original distribution contains errors, which may affect the Overlap and ROC AUC values for all metrics. This could explain why SiLVERScore's ROC AUC is slightly higher than those of other embedding-based metrics.

Metric	Overlap (%)	ROC AUC
BLEU-1	64.49	0.65
BLEU-2	71.50	0.63
BLEU-3	66.98	0.65
BLEU-4	69.47	0.63
ROUGE	67.45	0.67
BERTScore	78.19	0.55
BLEURT	81.31	0.47
SiLVERScore	83.49	0.60

Table 2: Overlap % and ROC AUC values for different metrics when comparing original and reordered sentence pairs. Embedding-based metrics maintain higher overlaps and lower ROC AUC values, suggesting that they capture semantic equivalences more effectively.

4.3 Which metric can evaluate multimodal and pragmatic aspects like prosody more effectively?

4.3.1 Motivation and Setup

Sign languages rely heavily on prosodic markers such as facial expressions, pauses, and intensity to

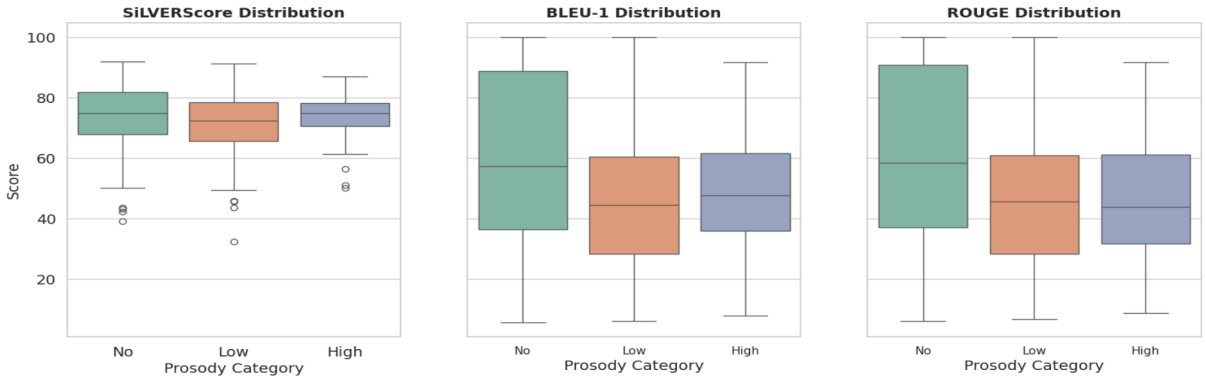

Figure 4: Box plots showing the distribution of SiLVERScore, BLEU-1, and ROUGE scores across three prosody intensity categories (No Intensity, Low Intensity, and High Intensity). While SiLVERScore remains stable across all categories, indicating robustness to prosodic variations, both BLEU-1 and ROUGE exhibit a noticeable decline in scores as prosody intensity increases. This drop suggests that BLEU-1 and ROUGE are sensitive to prosodically-rich sentences, which results in lower scores and higher variability in the High Intensity category.

convey meaning. Evaluating the robustness of metrics to prosodic variations is critical, as traditional back-translation-based methods often fail to capture such multimodal cues. We build on the work of Inan et al., 2022, which provided human-annotated token-level prosody intensities for the PHOENIX-14T dataset. These annotations classify tokens into three distinct prosodic levels: (i) no intensity: 0, indicating the absence of prosodic markers; (ii) low intensity: 1, reflecting a low degree of intensity markers; and (iii) high intensity: 2, representing high-degree intensity markers.

Sentence level prosody We define sentence intensity as the sum of the intensity levels of its tokens, $I = \sum_{i=1}^{n} t_i$, where t_i is the intensity of token i. Sentences are categorized into three prosody levels: No Intensity $I = 0$, Low Intensity $1 \leq I \leq 4$, and High Intensity $I \geq 5$.

Prosody level distribution The dataset exhibits the following distribution of sentences across these prosody categories: 328 sentences (51.09%) fall under No Intensity, 238 sentences (37.07%) under Low Intensity, and 76 sentences (11.84%) under High Intensity. This distribution indicates that the majority of sentences either lack prosodic markers or exhibit low levels of prosody, while highly expressive sentences are comparatively rare.

4.3.2 Distribution of Scores Across Prosody Categories

To analyze the impact of prosody on evaluation metrics, we categorized sentences based on the sentence-level intensity sums defined earlier. Figure 4 shows the distributions of SiLVERScore, BLEU-1, and ROUGE scores across the categories.

SiLVERScore Stability SiLVERScore remains consistent across the three prosody categories, showing minimal variation in median and interquartile range. This demonstrates that SiLVERScore effectively evaluates semantic alignment without being influenced by prosodic intensity.

BLEU-1 and ROUGE Sensitivity BLEU-1 and ROUGE scores decline with increasing prosody intensity, with median scores for High Intensity significantly lower than for No Intensity. This trend indicates that these metrics struggle with prosodically-rich sentences.

Score Variability Both BLEU-1 and ROUGE display higher variability in the High Intensity category, suggesting inconsistent performance in evaluating expressive signing.

4.4 Correlation with Prosodic Intensity

As shown in Table 3, traditional back-translation-based metrics (BLEU and ROUGE) exhibit significant negative correlations with prosody intensity (e.g., BLEU-4: -0.200, p = 3.31×10^{-7}), reflecting their vulnerability to prosodic variations. This behavior reflects the limitations of traditional metrics, which depend on surface-level text alignment and are vulnerable to information loss during back translation.

In contrast, SiLVERScore exhibits no significant correlation with prosody intensity (correlation: -0.004, p = 0.9277), indicating its robustness to

Metric	Correlation	p-value
BLEU-1	-0.160	< 0.01
BLEU-2	-0.178	< 0.01
BLEU-3	-0.191	< 0.01
BLEU-4	-0.200	< 0.01
ROUGE	-0.179	< 0.01
BERTScore	-0.144	< 0.01
BLEURT	-0.101	0.01
SiLVERScore	-0.004	0.93

Table 3: Pearson Correlation and p-value of metrics with sentence-level prosody intensity. BLEU and ROUGE exhibit significant negative correlations with prosody intensity, while SiLVERScore demonstrates no significant correlation.

prosodic variations. This robustness suggests SiLVERScore's ability to evaluate semantic alignment without being influenced by expressive elements.

5 The Generalization Problem

While evaluation metrics are expected to generalize across diverse datasets, this remains a significant challenge in sign language processing due to the limited size and diversity of available datasets. As highlighted by Jiang et al. (2024), one of the largest sign language dataset, SpreadtheSign, contains only 456,913 examples, which is orders of magnitude smaller than datasets in related domains (e.g., 400M examples for CLIP and 136M for VideoCLIP). In this section, we empirically demonstrate that even SignCLIP, the largest contrastive learning model to date, struggles with generalization at the token level.

5.1 Empirical Evidence of Limited Generalization

5.1.1 Token Level Generalization

We evaluated SignCLIP on ASL Citizen (Desai et al., 2024) and ASL Signs (Chow et al., 2023). The results show that SignCLIP's generalization capability is limited without fine-tuning. (Descriptions of these datasets can be found in Appendix E.)

Figure 5 illustrates the cosine similarity between video and text embeddings. Ideally, high similarity values should appear along the diagonal, indicating alignment between corresponding video-text pairs. Before fine-tuning, the heatmaps display low, diffuse similarity scores, indicating poor video-text alignment. Fine-tuning significantly improves alignment, indicating the necessity of dataset-specific adaptation. A similar trend is observed for ASL Signs (figures in Appendix B).

5.1.2 Sentence Level Generalization

We evaluated SignCLIP's sentence-level generalization on the WMTSLT Focus News Corpus (Mathias et al., 2022). (A description of this dataset is available in Appendix E.) Despite fine-tuning, SignCLIP struggles to achieve strong results (R@1 = 0.0436). Heatmaps (Figure 5) reveal diffuse patterns before fine-tuning and overfitting after, due to the dataset's limited size (9000 instances).

5.1.3 Token Level Language Specific Generalization

To investigate the effect of data size on generalization, we fine-tuned SignCLIP using combined training samples from ASL Signs and SemLex datasets. Despite this, SignCLIP fails to generalize effectively to ASL Citizen (R@5 = 0.0005). Even when training on all three datasets, the test set performance on ASL Citizen did not improve significantly. This suggests that dataset-specific characteristics influence performance even when substantial training data is available.

5.1.4 Representation Density

Ye et al., 2024 indentified a representation density problem, where the semantic visual representations of different sign gestures tend to be closely clustered together, making them hard to distinguish. The proposed contrastive learning strategy, SignCL, encourages the learning of discriminative feature representations. However, applying SignCL to our data yielded limited improvement in retrieval results (R@1 = 9.11E-05), compared to (R@1 = 3.04E-05) with vanilla contrastive learning.

5.1.5 Data Augmentation

Data augmentation is a commonly employed technique to improve model generalization, especially in domains with limited data. To this end, we experimented with several data augmentation strategies including: spatial 2D augmentation, temporal augmentation, and Gaussian noise on keypoints (Jiang et al., 2024). Results show negligible gains (R@1 = 0 with 2D-aug; 6.07E-05 with temporal augmentation), highlighting the limitations of conventional augmentation techniques in enhancing generalization. This suggests that limited dataset diversity and the complexity of visual sign representations cannot be fully addressed through conventional augmentation techniques alone.

Fine-tuned on	Tested on	SignCL	Data Aug	R @ 1	R @ 5	R @ 10
Token Level (§ 5.1.1)						
-	Citizen	-	-	0.0014	0.0061	0.0112
Citizen	Citizen	-	-	0.0639	0.2710	0.4392
Sentence Level (§ 5.1.2)						
WMTSLT	WMTSLT	-	-	0.0037	0.0175	0.0323
Token Level Language Specific (§ 5.1.3)						
Signs, SemLex	Citizen	-	-	3.04E-05	0.0005	0.0008
Citizen, Signs, SemLex	Citizen	-	-	0.0436	0.1764	0.2878
With SignCL (§ 5.1.4)						
Signs, SemLex	Citizen	✓	-	9.11E-05	0.0005	0.0009
With Data Augmentation (§ 5.1.5)						
Signs, SemLex	Citizen	-	2D-aug, Gaussian	0	0.0002	0.0006
Signs, SemLex	Citizen	-	Temporal	6.07E-05	9.11E-05	0.0003

Table 4: Text-to-Video Retrieval results and generalization across datasets. Results are shown for different fine-tuning datasets, test datasets, and configurations with or without data augmentation.

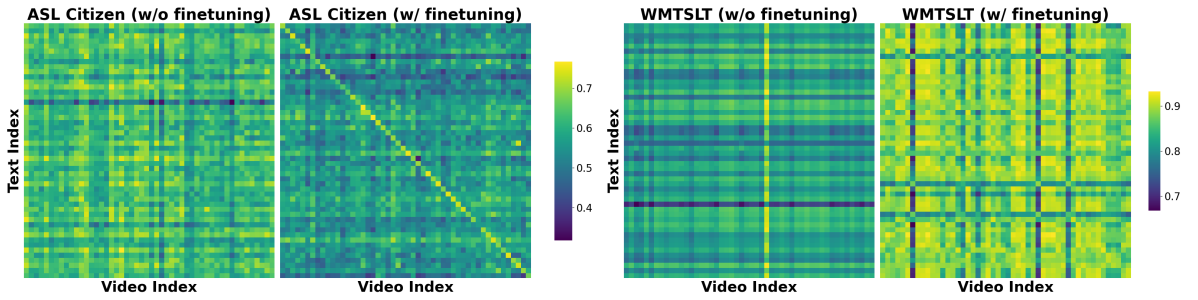

Figure 5: Heatmaps of SignCLIP embeddings cosine similarity scores for two datasets: ASL Citizen (token level) and WMTSLT (sentence level). **Left**: Finetuning increases alignment, as indicated by the clearer diagonal line. **Right**: After finetuning, the model appears to overfit, assigning high similarity scores to many pairs.

5.2 How SiLVERScore Addresses Generalization Challenges

Our findings from the experiments suggest the idea that, given the current constraints in data availability, tailoring metrics to specific datasets is necessary to create alignment between text and sign.

We proposed a dataset-specific evaluation metric designed to leverage the strengths of embedding-based methods while addressing the constraints of current sign language datasets. By optimizing for specific domains and datasets, we can achieve more reliable evaluations and better alignment with the linguistic and multimodal nature of sign language.

6 Conclusion

Through the introduction of SiLVERScore, we demonstrated the empirical strengths of embedding-based methods, including robustness to semantic variation, prosodic intensity, and a more holistic multimodal evaluation. Our results show that SiLVERScore can overcome limitations of traditional back-translation metrics.

SiLVERScore has the potential to reshape sign language evaluation standards by advancing accessibility for the Deaf community and promoting inclusivity in language technologies. Its robustness and semantic sensitivity make it well-suited for broader challenges in multimodal NLP, such as cross-lingual evaluation and integration with video generation models. To support open research and encourage further advancements, we release the code for SiLVERScore's analysis and computation.

Future efforts should integrate insights from computer graphics, such as improved modeling of spatial relationships and prosody in sign language, to further refine embedding-based methods. Incorporating richer multimodal features, including gesture dynamics and temporal coherence, could enhance the evaluation of expressive and context-dependent signing. Additionally, addressing the scarcity of diverse, large-scale datasets remains critical for improving model generalization.

7 Limitations

While the proposed metric, SiLVERScore, demonstrates strong empirical performance, this work has several limitations. One significant limitation is the absence of human evaluation. Although SiLVERScore shows clear advantages over traditional methods using back translation, it remains crucial to validate its alignment with human judgments. Human evaluators could provide insights into whether the metric effectively captures the semantic and linguistic aspects of generated sign language. Addressing this limitation will be a focus of future work.

Another limitation is the reliance on the PHOENIX-14T dataset, which centers on German Sign Language within the specific domain of weather forecasts. This narrow scope restricts the generalizability of SiLVERScore to other sign languages, domains, or datasets with broader semantic and linguistic diversity. Although we show how to adapt the embedding-based evaluation approach to a particular dataset, following similar data-specific adaptation procedures could allow the creation of comparable metrics for other sign language datasets as well.

The approach's reliance on translating textual annotations into English for alignment with CLIP embeddings poses challenges in multilingual scenarios. This reliance assumes that translation into English is both feasible and accurate, which may not hold in contexts involving less commonly studied languages with limited resources for translation.

Additionally, our current evaluation focuses on sentence-level retrieval, which overlooks the contextual dependencies and references made in prior sentences, as noted by Tanzer et al. (2024). Sign language often relies heavily on discourse-level context, and evaluating only at the sentence level may not fully capture these contexts.

Finally, while the results show that prosody does not degrade SiLVERScore's performance, this does not imply that the metric explicitly models prosody. Future research should investigate how to incorporate explicit prosodic modeling into evaluation metrics to better capture the expressive nuances of sign language.

Potential Risks. Adopting embedding-based metrics can inadvertently inherit biases, stereotypes, or inaccuracies from the underlying training data and models. If the pre-trained embeddings contain demographic, cultural, or linguistic biases, these may influence evaluations and potentially disadvantage certain signers or signing styles. Moreover, inaccuracies introduced at the text-annotation stage could propagate through the metric, reinforcing incorrect assessments. Finally, the metric's reliance on English textual embeddings and specific datasets may inadvertently privilege certain languages and cultures.

References

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)*.

Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, and Wenqiang Zhang. 2023. CiCo: Domain-Aware Sign Language Retrieval via Cross-Lingual Contrastive Learning . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19016–19026, Los Alamitos, CA, USA. IEEE Computer Society.

Ashley Chow, Glenn Cameron, Mark Sherwood, Phil Culliton, Sam Sepah, Sohier Dane, and Thad Starner. 2023. Google - isolated sign language recognition. `https://kaggle.com/competitions/asl-signs`. Kaggle.

Aashaka Desai, Lauren Berger, Fyodor O. Minakov, Vanessa Milan, Chinmay Singh, Kriston Pumphrey, Richard E. Ladner, Hal Daumé, Alex X. Lu, Naomi Caselli, and Danielle Bragg. 2024. Asl citizen: a community-sourced dataset for advancing isolated sign language recognition. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Lu Dong, Lipisha Chaudhary, Fei Xu, Xiao Wang, Mason Lary, and Ifeoma Nwogu. 2024. Signavatar: Sign language 3d motion reconstruction and generation. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10.

Mo Guan, Yan Wang, Guangkun Ma, Jiarui Liu, and Mingzu Sun. 2024. Multi-stream keypoint attention network for sign language recognition and translation. *ArXiv*, abs/2405.05672.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans

trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6629–6640, Red Hook, NY, USA. Curran Associates Inc.

Wencan Huang, Wenwen Pan, Zhou Zhao, and Qi Tian. 2021. Towards fast and high-quality sign language production. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3172–3181.

Eui Jun Hwang, Jung Ho Kim, Suk Min Cho, and Jong C Park. 2022. Non-autoregressive sign language production via knowledge distillation. *arXiv preprint arXiv:2208.06183*.

Eui Jun Hwang, Huije Lee, and Jong C. Park. 2024. A gloss-free sign language production with discrete representation. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–6.

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

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.

Jung-Ho Kim, Mathew Huerta-Enochian, Changyong Ko, and Du Hui Lee. 2024. SignBLEU: Automatic evaluation of multi-channel sign language translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Turin, Italy.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Scott K. Liddell. 2003. *Grammar, Gesture, and Meaning in American Sign Language*. Cambridge University Press.

Li Liu, Lufei Gao, Wentao Lei, Fengji Ma, Xiaotian Lin, and Jinting Wang. 2023. A survey on deep multi-modal learning for body language recognition and generation. *arXiv preprint arXiv:2308.08849*.

Müller Mathias, Ebling Sarah, Camgöz Necati Cihan, Jiang Zifan, Battisti Alessia, Moryossef Amit, Rios Annette, Bowden Richard, and Wong Ryan. 2022. Wmt-slt focusnews: Training data for the wmt shared task on sign language translation.

Amit Moryossef, Rotem Zilberman, and Ohad Langer. 2024. signwriting-evaluation: Effective sign language evaluation via signwriting. *arXiv preprint arXiv:2410.13668*.

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.

Amy Pu, Hyung Won Chung, Ankur P Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for mt. In *Proceedings of EMNLP*.

Wendy Sandler. 2012. The phonological organization of sign languages. *Language and Linguistics Compass*, 6(3):162–182. Epub 2012 Mar 2.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. Progressive transformers for end-to-end sign language production. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI*, page 687–705, Berlin, Heidelberg. Springer-Verlag.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021. Continuous 3d multi-channel sign language production via progressive transformers and mixture density networks. *Int. J. Comput. Vision*, 129(7):2113–2135.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of ACL*.

Garrett Tanzer, Maximus Shengelia, Ken Harrenstien, and David Uthus. 2024. Reconsidering sentence-level sign language translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6262–6287, Miami, Florida, USA. Association for Computational Linguistics.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. 2019. Fvd: A new metric for video generation.

Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2021. Read and attend: Temporal localisation in sign language videos. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16852–16861.

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

Pan Xie, Taiying Peng, Yao Du, and Qipeng Zhang. 2024. Sign Language Production with Latent Motion Transformer . In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3012–3022, Los Alamitos, CA, USA. IEEE Computer Society.

Jinhui Ye, Xing Wang, Wenxiang Jiao, Junwei Liang, and Hui Xiong. 2024. Improving gloss-free sign language translation by reducing representation density. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)*, 39:1 – 16.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Kernel Density Plots

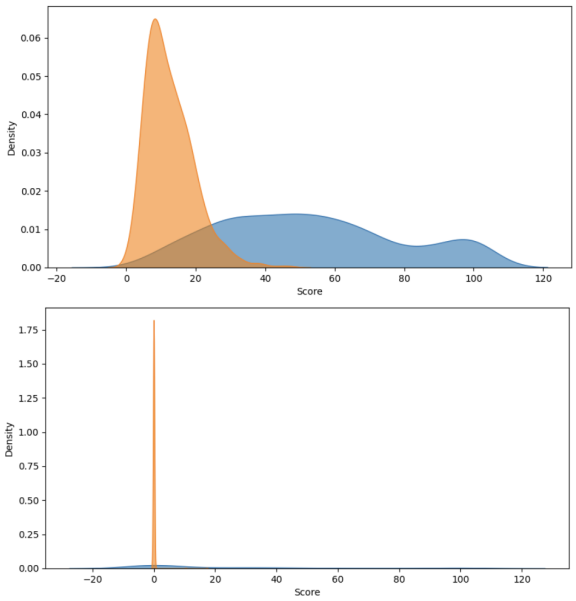

Figure 6: Kernel density plots for BLEU-1 (top) and BLEU-4 (bottom).

B Heatmaps

Figure 7 shows a heatmap of SignCLIP embedding cosine similarity scores for the ASL Signs dataset. A sharper diagonal pattern on the right indicates increased alignment between sign embeddings and their corresponding references.

C Training Details

Hardware and Compute All training and inference computations were performed on an NVIDIA A100 GPU with 80GB of GPU memory. The experiments were conducted on a Linux-based server environment equipped with 8 CPU cores.

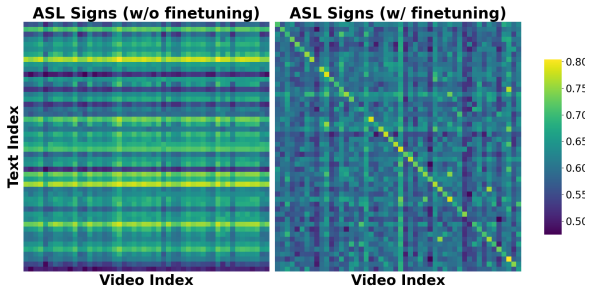

Figure 7: Heatmaps of SignCLIP embeddings cosine similarity scores for ASL Signs.

SignCLIP Fine-Tuning Fine-tuning the SignCLIP model on American Sign Language (ASL) datasets took the longest (approximately 4 hours). Fine-tuning was conducted using a batch size of 256, a maximum length of 64, Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.98$, and a gradient clipping norm of 2.0. Training involved 1,000,000 total updates and a warm-up phase over the first 122 updates. Up to 25 epochs were run with 1000 steps for monitoring.

Sign Language Translation (MSKA) Inference Inference with the MSKA model for sign language translation completed in under 10 minutes. Pretrained weights from Guan et al., 2024 were used without further fine-tuning.

Code and Configuration Files All code and configuration files necessary to reproduce the results (including model parameters, optimizer settings, and data preprocessing scripts) will be released via our GitHub repository upon paper acceptance.

D Prosody Boxplots

Figure 8 on the following page illustrates how various evaluation metrics distribute across sentences with different levels of prosodic intensity (No Intensity, Low Intensity, and High Intensity).

E Dataset Specifications

Table 5 summarizes the datasets used for our experiments in § 5, covering American Sign Language (ASL) and Swiss German Sign Language (DSGS).

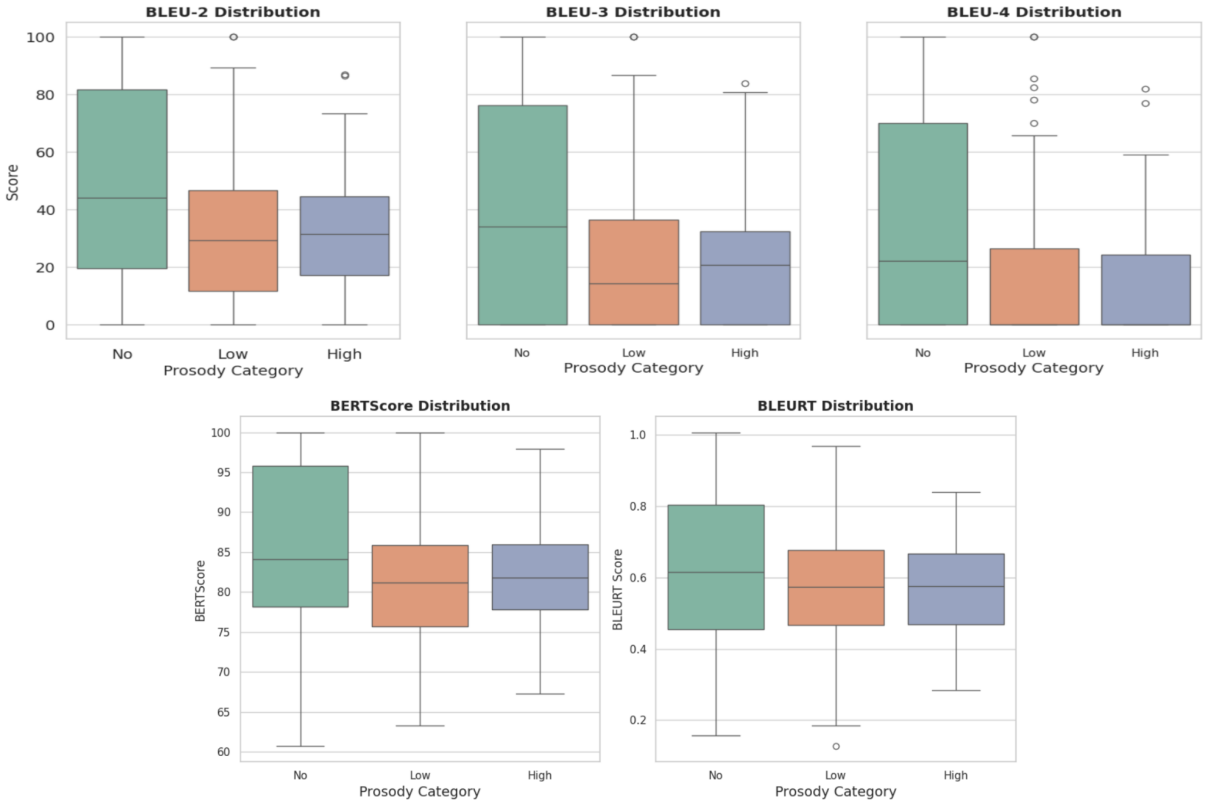

Figure 8: Box plots showing the distribution of BLEU-2, BLEU-3, BLEU-4, BERTScore, and BLEURT scores across three prosody intensity categories (No Intensity, Low Intensity, and High Intensity). Traditional back-translation metrics (BLEU) and embedding-based metrics (BERTScore and BLEURT) show a decline in scores with increasing prosody intensity

Dataset	Language	Level	# of Samples (Train/Val/Test)	# of Signers
ASL Signs	ASL	Token Level	85,031 / 4,723 / 4,723	100+ Signers
SemLex	ASL	Token Level	51,029 / 18,025 / 15,514	119 deaf signers
ASL Citizen	ASL	Token Level	40,154 / 10,304 / 32,941	52 deaf/hard-of-hearing
WMTSLT	DSGS	Sentence Level	9172 / 470 / 494*	12 deaf signers

Table 5: Overview of the datasets used in our evaluations. For the WMTSLT dataset, the train/validation/test split was generated by the authors, as the original dataset provided by the challenge did not include a predefined test set.