Beyond a Single Reference: Training and Evaluation with Paraphrases in Sign Language Translation

Anonymous ACL submission

Abstract

Sign Language Translation (SLT) corpora typically pair each signed utterance with a single written reference, despite the non-isomorphic relationship between signed and spoken languages that admits multiple valid translations. This constrains both training and n-gram-based evaluation. We investigate using LLMs to generate paraphrased references for SLT, comparing generation strategies via an adapted ParaScore metric and studying their effect on the pose-based T5 Baseline and Uni-Sign models trained on YouTubeASL and How2Sign. We find that naively incorporating paraphrases during fine-tuning does not improve translation quality, however, using paraphrases during evaluation yields scores that align more closely with human judgments. We formalize this as $\text{BLEU}_{\text{para}}$, a multi-reference extension of BLEU, and show via human evaluation that it correlates more strongly with perceived quality than standard BLEU — particularly on outlier translations. We release all paraphrases and code.

1 Introduction

Automatic Sign Language Translation (SLT) has recently made rapid progress due to advances in Vision Language Models and the availability of large-scale video–text corpora. Despite these developments, a persistent limitation in current SLT datasets is the scarcity of translation variability: each sign language utterance is typically paired with a single weakly aligned written-language reference translation. This stands in contrast to spoken language machine translation, where corpora frequently provide multiple reference translations or explicitly annotate translation alternatives, capturing the natural diversity of phrasing, register, and information structure that exists in human language. Such variation is crucial for both training and evaluation, where metrics such as BLEU are known to be sensitive to the availability of reference translations due to their n-gram matching nature.

Sign languages pose an even stronger requirement for translation diversity. The mapping between visual expression and written language is highly non-isomorphic: a single sign language sentence can be translated into several equally valid spoken language realizations, differing in word order, degree of explicitness, and syntactic structure. Relying on a single reference translation therefore risks over-constraining model training and underestimating translation quality during evaluation.

In this work, we explore whether Large Language Models (LLMs) can be used to systematically enrich SLT corpora with automatically generated paraphrases of written language translations. Our goal is twofold: (1) to study whether synthetic generation of translation variants (i.e. paraphrasing) can improve the robustness of sequence-to-sequence SLT models, and (2) to examine how these paraphrases affect evaluation, comparing BLEU scores against human evaluation.

By viewing paraphrases as alternative translations, analogous to multi-reference machine translation setups, we aim to bridge principles from classical natural language processing (NLP) with the specific challenges of SLT.

2 Related work

Sign Language Translation Recent progress in SLT have been driven by increasingly powerful multimodal and transformer-based architectures. A commonly used baseline architecture consists of a keypoint extractor (e.g. MMPose (Sengupta et al., 2020) or MediaPipe (Lugaresi et al., 2019)), followed by transformer (Vaswani et al., 2017). This approach became an effective way to perform experiments and provide benchmark results for new datasets (Uthus et al., 2023; Camgöz et al., 2021; Tanzer and Zhang, 2025; Tanzer, 2025; Al-

hejab et al., 2025). Recent state-of-the-art methods extend this approach by adding additional modules. GloFE (Lin et al., 2023) employs visual backbone based on graph convolutional networks (GCN) to more effectively process pose features, followed by a transformer encoder–decoder architecture that learns robust gloss-level representations through contrastive learning. Uni-Sign (Li et al., 2025) similarly uses GCN for pose feature encoding and further incorporates a fusion module that integrates RGB information in cases where keypoint extraction is failing. SignBERT+ (Hu et al., 2023) extends BERT-style pretraining (Devlin et al., 2019) to sign language data, learning contextualized representations that capture temporal dependencies within signing sequences and can be transferred to downstream tasks such as isolated and continuous sign language recognition. Another recent direction moves beyond gloss-based supervision toward sentence-level semantic supervision. Hamidullah et al. (Hamidullah et al., 2024) propose using target-sentence embeddings as an intermediate training signal in place of manual glosses, enabling a gloss-free and multilingual SLT setup. Evaluated on PHOENIX-2014T and How2Sign, their approach outperforms prior gloss-free systems and achieves state-of-the-art results on How2Sign when no additional SLT data is used for pretraining. LITFIC (Jang et al., 2025) leverages a LLaMA-based (Touvron et al., 2023) large language model that incorporates contextual and meta-information, including pseudo-gloss cues, to improve the translation quality.

Paraphrasing Paraphrasing - generating semantically equivalent expressions with varied surface forms - has been extensively studied for both its theoretical properties and practical applications in NLP. Bhagat & Hovy (Bhagat and Hovy, 2013) established that quality paraphrases balance semantic preservation against lexical and syntactic diversity, a tension later formalized in evaluation metrics like ParaScore (Shen et al., 2022), which combines BERTScore-based (Zhang et al., 2020) semantic similarity with explicit diversity modeling. Neural paraphrase generation has evolved from seq2seq models to transformer-based approaches fine-tuned on large-scale datasets like ParaNMT-50M (Iyyer et al., 2018). In machine translation, paraphrasing serves as a key data augmentation strategy, particularly for low-resource settings where back-translation (Sennrich et al., 2016) and lexical substitution have yielded substantial BLEU improvements. Paraphrasing is particularly relevant for SLT: the structural divergence between signed and written languages means a single signed utterance often admits multiple valid translations, yet datasets typically provide only one. Recent work further shows that this is not only a data issue, but also an evaluation issue: Yazdani et al. (Yazdani et al., 2025) find that lexical overlap metrics such as BLEU, chrF, and ROUGE are highly sensitive to paraphrased SLT outputs, with BLEU dropping sharply despite preserved meaning, while multi-reference evaluation better rewards semantically equivalent translations. These findings directly motivate our use of paraphrased references for evaluating SLT, while our work differs by generating paraphrases for dataset references and validating the resulting $\text{BLEU}_{\text{para}}$ score against human judgments.

Data In this research, we focus on American Sign Language, namely by using two widespread datasets - YoutubeASL (Uthus et al., 2023) and How2Sign (Duarte et al., 2021). YouTubeASL is a large-scale dataset of American Sign Language (ASL) collected from publicly available YouTube videos. It contains over 610,000 video samples (equivalent to about 1,000 hours) and features 2,500 unique signers from diverse backgrounds. Videos cover various domains, such as education, personal vlogs, and conversational ASL, making the dataset representative of a wide range of signing contexts.

How2Sign is a smaller but well-annotated dataset, originally developed to study sign translation under multiple angles/views. It contains 35,000 signing samples, amounting to approximately 80 hours of ASL content, produced by roughly 11 signers in a controlled environment with stable lighting and fixed backgrounds. Although the dataset's primary focus is translating from spoken English to ASL, it is widely adopted for other tasks in the sign language processing community, including sign language translation. One challenge with using How2Sign lies in the presence of translationese effects, a phenomenon where translated text can sound artificial or lack natural linguistic flow (Graham et al., 2020). Despite this limitation, How2Sign serves as an important benchmark for evaluating model performance. Its controlled nature provides consistency that complements the more variable, in-the-wild data from YouTubeASL.

3 Methods

3.1 Sign Language Translation

For our experiments, we adopt an existing SLT framework that uses a modified T5 encoder–decoder transformer to process pose-based inputs and includes comprehensive data preprocessing, such as keypoint extraction, normalization, interpolation, and augmentation (Zelezny et al., 2025). We adopt this existing setup because it provides already ready-to-use codebase and preprocessed YouTubeASL dataset (Zelezny et al., 2024), including extracted keypoints and corresponding annotations. To evaluate the generalizability of our proposed approach beyond a single architecture, we additionally conduct experiments with the Uni-Sign (Li et al., 2025) model. In this case, we reuse the preprocessing pipeline and pose-based inputs prepared within the T5 baseline framework, allowing both models to be trained and evaluated on identical data representations and under consistent experimental conditions.

3.2 Paraphrasing and ParaScore

Paraphrasing plays a vital role in modern natural language processing and machine translation, making the quality assessment of paraphrases essential for these applications. In the context of SLT, where the mapping between visual and written modalities admits multiple valid realizations, paraphrases can serve as synthetic valid alternative references. However, not all automatically generated paraphrases will be equally useful: a paraphrase that diverges too far semantically fails to represent the source meaning, while one that is too lexically similar to the original provides little additional signal. A successful paraphrase must balance two critical criteria: semantic similarity and lexical divergence. To evaluate this balance, we employ an adapted version of the ParaScore metric (Shen et al., 2022), which effectively combines semantic preservation with linguistic diversity through the use of both language models and traditional linguistic measures.

Unlike most existing metrics that struggle to adequately balance semantic similarity and lexical divergence, this approach implements a max function for the semantic similarity component between an input sequence and its candidate rephrasement. The semantic similarity component utilizes BERTScore (Zhang et al., 2020), which employs the following process: it first computes contextual embeddings for each token in both sentences, then calculates cosine similarities between tokens in the two sentences, and finally employs a matching strategy to compute precision, recall, and F1 score.

For measuring lexical divergence, we implement the Levenshtein distance (Levenshtein, 1965), which calculates the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one word into another. This distance is normalized by dividing by the length of the longer string, resulting in a value between 0 and 1, where 0 indicates identical strings and 1 represents completely different strings.

The metric incorporates two hyperparameters: γ and ω. γ, set to a default value of 0.35, serves two purposes: it caps the maximum divergence score and introduces non-linearity to the scaling of the divergence score. This means that edit distances above 35% of the longer string's length are treated equivalently. ω, with a default value of 0.5, is used in the final ParaScore calculation to weight the divergence. Formally, given an input x and a candidate paraphrase $\hat{x}$, the metric is defined as:

$$\text{ParaScore}(x, \hat{x}) = \frac{\text{BERTScore}(x,\hat{x}) + \omega \cdot \min(\text{NLD}(x,\hat{x}), \gamma)}{1 + (\omega \cdot \gamma)} \quad (1)$$

where NLD is the Normalized Levenshtein Distance. With these parameters, the similarity score can contribute up to 1 to the final score, while the divergence score can contribute up to 0.175 (i.e., 0.5×0.35). Although this creates a maximum possible score of 1.175, we normalize the final score by dividing by this maximum (represented by the denominator) to achieve a standardized scale.

4 Comparison of Paraphrasing Methods

In this section we compare paraphrasing models under a single, controlled generation pipeline. For each reference English sentence x in our SLT corpora, we ask a candidate LLM to produce $K{=}5$ meaning-preserving rewrites $\{\hat{x}_k\}_{k=1}^{K}$ using the same system instruction across models: `You are a helpful assistant that rephrases a given sentence in K ways, each on its own line. Try to be semantically consistent and output nothing else than these sentences. Sentence: <x> Paraphrases:`. To keep outputs comparable, we use consistent decoding (sampling; temperature $= 0.7$, top-$p{=}0.95$). We then normalize generations by stripping list markers/bullets and occasional boilerplate text (e.g., "Here are ...");

sentences outside 4–30 words are skipped, and any generation that does not yield exactly K lines is treated as missing. The resulting per-model paraphrase sets are scored with ParaScore and used for the analyses in FiguresFigures 1, 2, 3 and the subsequent experiments.

Through careful manual evaluation of paraphrases at various percentile levels (0.25, 0.50, 0.75), we established a quality threshold of 0.7. This threshold serves as a reliable indicator for distinguishing high-quality paraphrases from lower-quality ones.

For paraphrasing ASL translations, we use a wide range of current Large Language Models. Their results are compared in Figure 1.

Because SLT is currently done mostly at sentence-level, the given sentence is usually a few seconds clip, extracted from a longer video, lacking context, which was proven particularly problematic as Sign Languages are very context-dependent (Jang et al., 2025).

As a separate experiment, we augment the paraphrasing prompt with lightweight *video-level textual context*. For each target sentence, we add the preceding reference sentence(s) from the same video clip, and reset the context when moving to a new video. This context is provided as part of the instruction while keeping the target sentence unchanged, allowing us to isolate whether discourse information (co-reference, ellipsis, topic continuity) improves meaning-preserving paraphrases compared to sentence-only prompting. However, in our experimental setting the ParaScore is worse for the paraphrases generated with context (see Figure 2). This is something we are planning to explore further in our future work.

Another separate experimental setting tackles the use of prompting to gain multiple paraphrases of the same reference translation sentence. We can either use **sequential** (ask for n paraphrases at once in a single prompt) or **iterative** prompting (ask for a paraphrase n times). Sequential prompting proved to generate paraphrases with a little higher average ParaScore (see Figure 3) and for a lower computation cost.

In our comparison, we applied our adapted ParaScore metric to each dataset, analyzing both individual paraphrases and the average scores for each caption. The analysis revealed that GPT-4o-mini paraphrasing achieved the highest average ParaScore; therefore, its generated paraphrase sets for YouTubeASL and How2Sign are used in the subsequent training and evaluation experiments. Notably, we observed that adding context to both sequential and iterative prompting in LLaMA led to significant performance degradation.

5 Paraphrases for Training and Evaluation

In this section, we investigate the impact of paraphrases on SLT with two main objectives: (1) To examine whether incorporating paraphrased target sentences during fine-tuning improves SLT performance. (2) To assess whether using paraphrases can lead to improved automatic evaluation quality.

Different paraphrase-based fine-tuning and evaluation strategies are compared in a controlled experimental setting. In this section, paraphrases are introduced only during the How2Sign fine-tuning stage; pretraining on YouTubeASL always uses the canonical references. Evaluation using paraphrased references is particularly interesting, as higher alignment with human judgment would indicate their potential usefulness; the analysis of human judgments is discussed in Section 6.

5.1 Experimental Setup

We follow the default training configurations and preprocessing settings recommended in the original works of the adopted methods (see Section 3.1). Experiments on the YouTubeASL dataset use the provided 90:10 train–validation split with pre-extracted pose keypoints, while for How2Sign we use the original dataset split. To ensure fair comparison across architectures, both the T5 baseline and Uni-Sign models are trained and evaluated using the same preprocessed pose representations. Model-specific training hyperparameters are summarized in Table 1.

Model	T5 Baseline	Uni-Sign
batch size	256	128
learning rate	4e-4	3e-4
lr scheduler	constant	cosine
normalization	SignSpace	none
warmup steps	5000	0
training epochs	20	

Table 1: Hyperparameters used for training. The same hyperparameters were used for both YouTubeASL and How2Sign datasets.

For evaluation, we select the best checkpoint on the dev set. We report test set BLEU scores computed using the `SacreBLEU` library

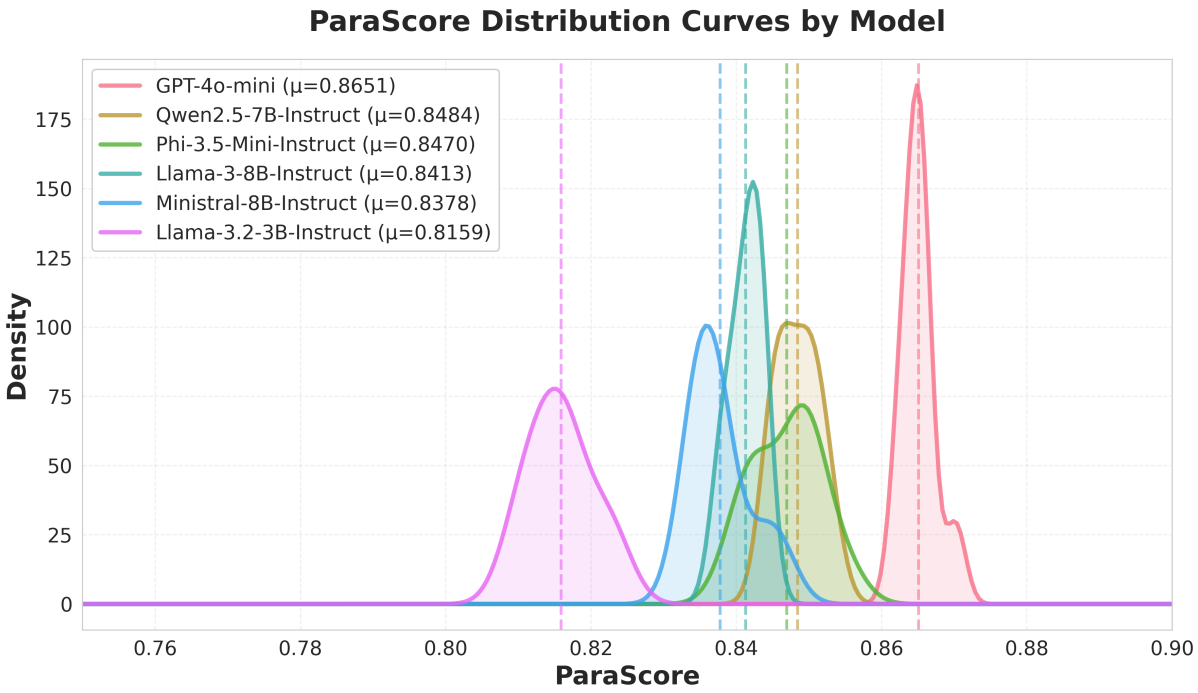

Figure 1: **ParaScore distributions for paraphrases generated by different LLMs.** Kernel density estimates over all generated paraphrases (same prompt and decoding for all models). Dashed vertical lines mark per-model means (μ; shown in the legend). Higher scores indicate paraphrases that better preserve meaning while avoiding near-copies, enabling a direct quality comparison across paraphrasing models.

v2.4.3[1], ROUGE-L scores from `rouge-score` library v0.1.2[2], and BLEURT scores using the `BLEURT-20` checkpoint from the official repository[3].

5.2 Experiments

First, the models are pretrained on the YouTube ASL dataset without paraphrases. They are then fine-tuned on the How2Sign dataset using different types of paraphrase-based training strategies. The resulting model checkpoints are then evaluated using the How2Sign test set. To ensure consistency of results, each experiment is run three times with different seeds, and all reported results correspond to the average performance across these runs. Complete results are provided in the supplementary material.

We consider three fine-tuning configurations. First, we train the models using only the canonical target sentence without paraphrases, denoted as *No par.*. Second, we train the models where the target sentence is randomly sampled from up to six alternatives (one canonical reference and five paraphrases) for each training instance, denoted as *Random*. Third, we compute the loss for all available paraphrases and backpropagate the gradients from the paraphrase yielding the minimum loss for a given instance, denoted as *Min loss*.

In the paraphrase-based evaluation setting, BLEU is computed as $\text{BLEU}_{\text{para}}$ using a multi-reference BLEU with all six available references (one canonical reference and five paraphrases) using `SacreBLEU` library. For BLEURT and ROUGE-L, we compute the score against each available reference independently and report the maximum score per test instance. All metrics are then averaged over the test set.

As shown in Table 2, the effect of training with paraphrased targets varies across models. For the T5 baseline, fine-tuning without paraphrases consistently yields better performance than training with paraphrased targets. In contrast, for Uni-Sign, the min-loss setting is comparable to training without paraphrases. This suggests that the impact of paraphrased supervision is model-dependent, and that selecting higher-quality paraphrases can mitigate the ambiguity introduced by multiple target variants. As expected, evaluation scores computed using multiple references are generally higher than single-reference settings. A more detailed com-

[1] https://pypi.org/project/sacrebleu/2.4.3/
[2] https://pypi.org/project/rouge-score/0.1.2/
[3] https://github.com/google-research/bleurt

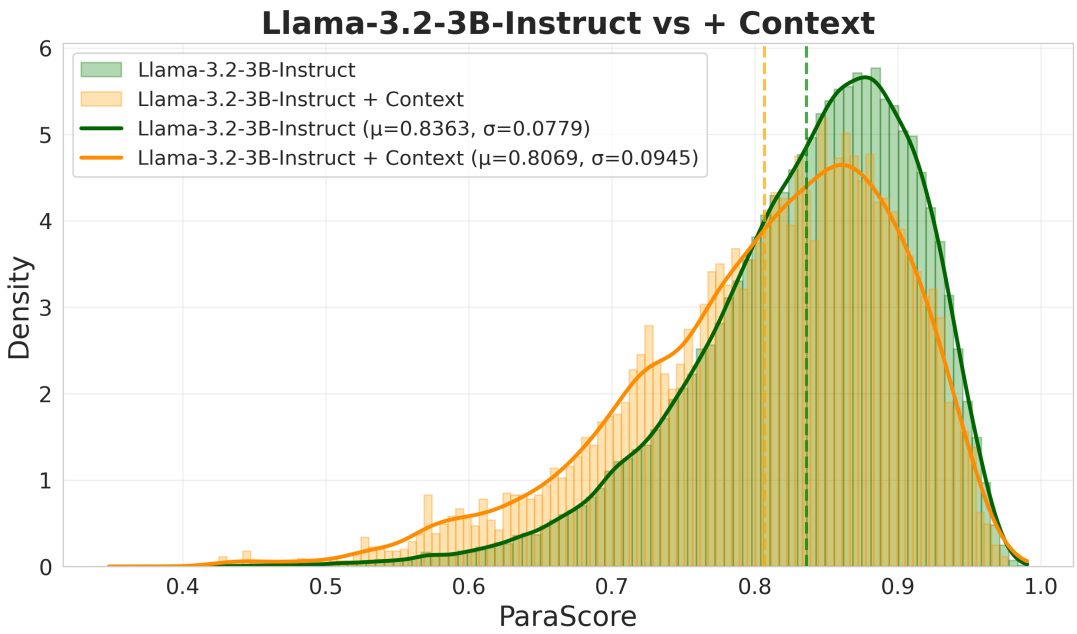

Figure 2: **Effect of adding video-level context to the paraphrasing prompt.** ParaScore distributions for Llama-3.2-3B-Instruct with sentence-only prompting versus prompting augmented with short preceding context from the same video clip. Dashed lines denote means (μ); providing context shifts the distribution left and increases variance, suggesting that extra discourse information can encourage looser rewrites that more often drift from the original meaning under our ParaScore criterion.

parison with human judgments is provided in the following section.

Table 2 also shows that paraphrased references affect metrics differently. For the T5 baseline, BLEU-4 increases from 7.46 to 8.86, corresponding to an approximately 19% relative gain; for Uni-Sign, it increases from 11.68 to 13.39 under the no-paraphrase fine-tuning setting whereas BLEURT increases only marginally from 42.7 to 44.5 and 45.3 to 46.7, about a 3% relative gain. This asymmetry is expected: embedding-based metrics such as BLEURT are designed to be less sensitive to surface-form variation, while n-gram metrics such as BLEU are directly affected by synonym choice and word order. Thus, $\text{BLEU}_{\text{para}}$ primarily addresses a limitation of reference-matching metrics, which remain dominant in SLT evaluation, rather than replacing semantic metrics such as BLEURT.

6 Human Evaluation

To address prior findings (Jang et al., 2025), which show concerning limitations of standardized metrics in SLT, usually focused on WER and translation length, we propose the $\text{BLEU}_{\text{para}}$ evaluation protocol, which extends BLEU-4 (Papineni et al., 2002) by evaluating each prediction against six available references: one canonical reference and five paraphrases. In contrast to the single-reference setting, these references are provided jointly to the multi-reference BLEU computation using `SacreBLEU`, following the evaluation setup described in Section 5. This partially alleviates penalties due to synonym choice and word-order variation while preserving the standard BLEU formulation. To enable its use, we publish the LLM-generated paraphrased versions of How2Sign (Duarte et al., 2021) and YouTubeASL (Uthus et al., 2023) alongside this paper .

To prove that $\text{BLEU}_{\text{para}}$ improves SLT evaluation over its non-paraphrased counterpart, we have conducted a Human Evaluation experiment, using the exact same protocol as (Jang et al., 2025). Due to the significant annotation cost and time requirements of human evaluation, this experiment was conducted only for the T5 Baseline. In addition to reporting correlations over all items, we explicitly probe *extreme* cases where reference-matching metrics are typically the most brittle: we select the 48 examples whose canonical-reference BLEU is either very low (< 5) or very high (> 15). This subset captures both clear failures and near-verbatim matches, and highlights whether $\text{BLEU}_{\text{para}}$ better tracks human judgments in outlier translations.

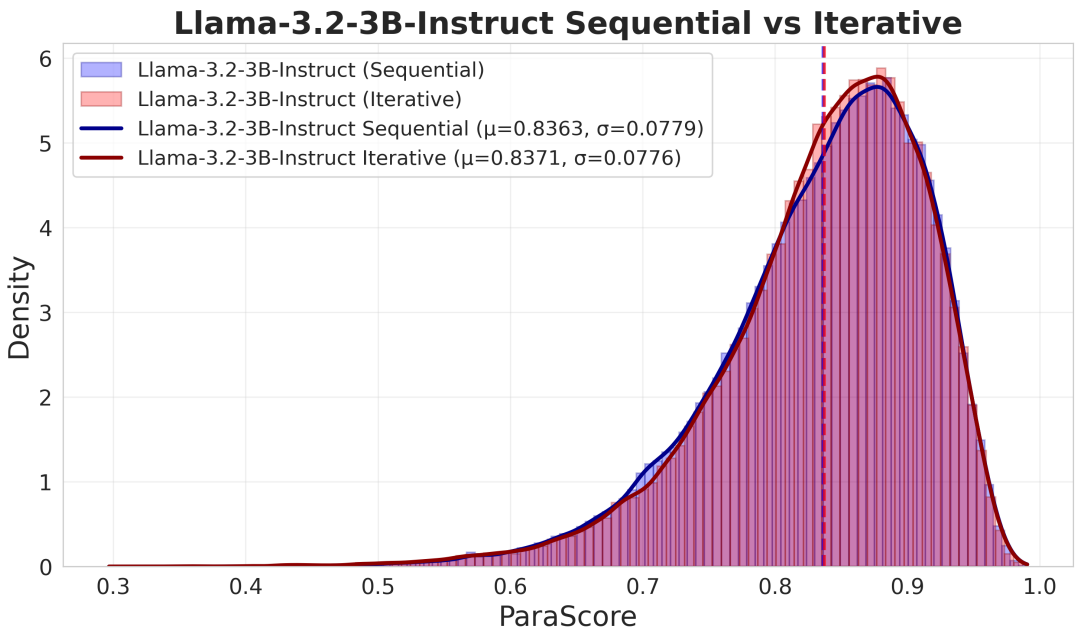

Figure 3: **Sequential vs. iterative paraphrase generation.** ParaScore distributions for Llama-3.2-3B-Instruct when producing five paraphrases in a single call (*sequential*) versus generating them one-by-one with feedback from earlier outputs (*iterative*). Histograms and density curves largely overlap, and the mean/variance (dashed lines; μ, σ in legend) are nearly identical, indicating that iterative prompting does not materially change paraphrase quality in our setting.

Par. mode	BLEU-4	BLEURT	Rouge-L
Evaluation with canonical reference only			
No par.	**7.46**	42.63	**28.94**
Random	6.17	**42.74**	26.50
Min loss	6.49	42.64	27.48
Evaluation with paraphrases			
No par.	**8.86**	44.10	**30.91**
Random	7.95	**44.48**	29.64
Min loss	8.00	44.28	30.09

(a) T5 baseline evaluation

Par. mode	BLEU-4	BLEURT	Rouge-L
Evaluation with canonical reference only			
No par.	**11.68**	**45.30**	33.17
Random	9.46	44.55	29.78
Min loss	11.58	45.13	**32.59**
Evaluation with paraphrases			
No par.	13.39	46.64	**35.10**
Random	11.59	46.14	32.60
Min loss	**13.52**	**46.67**	34.83

(b) Uni-Sign evaluation

Table 2: Performance of different paraphrase-based fine-tuning strategies on the How2Sign test set. Results are reported for canonical-reference evaluation and paraphrase-aware evaluation ($\text{BLEU}_{\text{para}}$). `No par.` uses only canonical targets, `Random` samples from available paraphrases during fine-tuning, and `Min loss` selects the target with the minimum fine-tuning loss.

The results in Table 3 show that human perception of model capabilities correlates better with $\text{BLEU}_{\text{para}}$, especially in outlier cases (48 translations with BLEU < 5 or > 15).

7 Discussion

Our comparison of paraphrasing methods reveals substantial variability among individual LLMs in their ability to generate high-quality paraphrases for synthetic translation data. Among the evaluated models, GPT-4o outperforms the other methods by a considerable margin. We assume the trend to be improving, with the expectation that commercial models will be slightly more capable than publicly available models. It is therefore likely that the results achieved by GPT-4o in our experiments may already be surpassed by newer versions of commercial models, such as GPT-5 or Gemini 3. Nevertheless, because these experiments are finan-

Metric	Pearson r	Spearman ρ	ρ (extremes)
BLEU	0.688	0.657	0.578
$\text{BLEU}_{\text{para}}$	**0.697**	**0.685**	**0.659**

Table 3: Correlation between mean human ratings (0–5; 70 translations, 6 annotators) and automatic metrics. The "extremes" subset contains 48 examples with canonical-reference BLEU < 5 or > 15.

cially demanding and the primary goal of this paper is not to achieve state-of-the-art results, but rather to explore a general concept, we leave additional experiments for future research.

Incorporating paraphrases directly as data augmentation during model fine-tuning does not yield measurable performance gains, suggesting that naive output-level augmentation is ineffective. This phenomenon is consistent for both tested architectures. While introducing multiple paraphrase variants into the loss computation could potentially improve learning, this approach significantly increases training time in the current setup. More selective strategies, such as hard example mining, may therefore be a more promising direction for future research.

Prior work has demonstrated that current standardized metrics, such as BLEU and ROUGE, rely heavily on word-level matching strategies, which correlate with human perception only weakly. To address this limitation, we propose a novel evaluation protocol $\text{BLEU}_{\text{para}}$. Human evaluation confirms that the correlation is improved by employing this instead of the standard BLEU. Additionally, $\text{BLEU}_{\text{para}}$ partially alleviates issues related to synonyms and word order. Its main disadvantage is the necessity to have paraphrases for each test sample. To reduce this burden and support reproducibility, we publish all paraphrases, including the codes for their creation, in our GitHub repository .

8 Conclusion

This paper experiments with the use of paraphrases for training and evaluation. We first compare the capabilities of different large language models in the task of paraphrase creation. Then we utilize these paraphrases for both fine-tuning and evaluation. While the training with paraphrases, using basic ideas for its incorporation, does not bring any measurable improvements, we would like to explore more complex protocols in our future research.

On the other hand, the evaluation using paraphrases provides results better correlated with human perception of model capabilities. To reflect the usage of paraphrases during the evaluation step, we propose a novel evaluation protocol $\text{BLEU}_{\text{para}}$. We argue that this protocol is more suitable for the evaluation of methods for SLT than the standard metrics like BLEU or ROUGE. We validate these findings on two different architectures, T5 Baseline and Uni-Sign, demonstrating that the observed trends are not tied to a single model and suggesting the model-invariant nature of the proposed evaluation approach. All the codes and paraphrases are publicly available in our GitHub repository .

In our future work, we would like to explore different protocols for paraphrase creation more in depth, especially the utilization of context. Additionally, we would like to test if $\text{BLEU}_{\text{para}}$ generalizes well also for other sign languages other than American Sign Language.

Limitations

The main limitation of this research is scope and reproducibility. Our experiments are confined to a small set of datasets and one translation setting, so the findings may not transfer to other sign languages, domains, or annotation styles. Results are also sensitive to the choice of paraphrasing model, decoding settings, and prompts; in particular, using proprietary LLMs and stochastic sampling can make exact reproduction difficult and may change over time as providers update models. Finally, compute and cost constraints limited the number of paraphrases per reference, the breadth of hyperparameter sweeps, and the scale of human evaluation, so some effects (especially small gains between strong models) may be underpowered and should be revalidated on larger samples.

Our work is limited by the technical properties of sign language video data and its alignment with translations. Variations in resolution, frame rate, motion blur, camera framing, lighting, occlusions, and background contrast can reduce the visibility of important manual and non-manual cues. Video segments are often temporally incomplete or misaligned with their translations, lacking natural onset and offset phases, which introduces noise and complicates training and evaluation. These factors affect both the reliability of model learning and the interpretability of automatic metrics, including experiments with paraphrased targets and evaluation

references. In particular, the negative fine-tuning result depends on our selected paraphrasing model, prompt, ParaScore threshold, and the resulting balance between semantic fidelity and lexical variation; different paraphrase distributions may lead to different training behavior.

References

Ali Alhejab, Tomas Zelezny, Lamya Alkanhal, Ivan Gruber, Yazeed Alharbi, Jakub Straka, Vaclav Javorek, Marek Hruz, Badriah Alkalifah, and Ahmed Ali. 2025. Saudi sign language translation using t5. In *International Conference on Speech and Computer*, pages 331–343.

Rahul Bhagat and Eduard Hovy. 2013. Squibs: What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Necati Cihan Camgöz, Ben Saunders, Guillaume Rochette, Marco Giovanelli, Giacomo Inches, Robin Nachtrab-Ribback, and Richard Bowden. 2021. Content4all open research sign language translation datasets. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–5.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

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

Hezhen Hu, Weichao Zhao, Wengang Zhou, and Houqiang Li. 2023. Signbert+: Hand-model-aware self-supervised pre-training for sign language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11221–11239.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Youngjoon Jang, Haran Raajesh, Liliane Momeni, Gül Varol, and Andrew Zisserman. 2025. Lost in translation, found in context: Sign language translation with contextual cues. *arXiv preprint arXiv:2501.09754*.

Vladimir I. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR*, 163(4):845–848.

Zecheng Li, Wengang Zhou, Weichao Zhao, Kepeng Wu, Hezhen Hu, and Houqiang Li. 2025. Uni-sign: Toward unified sign language understanding at scale. In *The Thirteenth International Conference on Learning Representations*.

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Yi Yang, and 1 others. 2023. Gloss-free end-to-end sign language translation. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, and 1 others. 2019. Mediapipe: A framework for perceiving and processing reality. In *Third workshop on computer vision for AR/VR at IEEE computer vision and pattern recognition (CVPR)*, volume 2019.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Arindam Sengupta, Feng Jin, Renyuan Zhang, and Siyang Cao. 2020. mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns. *IEEE sensors journal*, 20(17):10032–10044.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Lingfeng Shen, Lemao Liu, Haiyun Jiang, and Shuming Shi. 2022. On the evaluation metrics for paraphrase generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3178–3190, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Garrett Tanzer. 2025. FLEURS-ASL: Including American Sign Language in massively multilingual multitask evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6167–6191, Albuquerque, New Mexico. Association for Computational Linguistics.

Garrett Tanzer and Biao Zhang. 2025. Youtube-SL-25: A large-scale, open-domain multilingual sign language parallel corpus. In *The Thirteenth International Conference on Learning Representations*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Dave Uthus, Garrett Tanzer, and Manfred Georg. 2023. Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus. In *Advances in Neural Information Processing Systems*, volume 36, pages 29029–29047. Curran Associates, Inc.

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

Shakib Yazdani, Yasser Hamidullah, Cristina España-Bonet, Eleftherios Avramidis, and Josef van Genabith. 2025. A critical study of automatic evaluation in sign language translation. *Preprint*, arXiv:2510.25434.

Tomas Zelezny, Marek Hruz, Jaub Straka, and Shester Gueuwou. 2024. YouTube-ASL clip keypoint dataset. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Tomas Zelezny, Jakub Straka, Vaclav Javorek, Ondrej Valach, Marek Hruz, and Ivan Gruber. 2025. Exploring pose-based sign language translation: Ablation studies and attention insights. *arXiv preprint arXiv:2507.01532*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

A Supplementary material

A.1 SLT Evaluation Details

In order to preserve consistency of our experiments, we run each of our experiments with three different seeds, as discussed in Section 5. Detailed results for all of our runs and all metrics we evaluated on are reported in Table 4.

T5 Baseline							
Paraphrase mode	Seed	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEURT	Rouge-L
Evaluation without paraphrases							
No paraphrases	42	28.34	16.38	10.76	7.44	42.56	28.88
No paraphrases	0	27.16	15.90	10.59	7.45	42.94	**29.37**
No paraphrases	1	**28.69**	**16.50**	**10.83**	**7.48**	42.39	28.58
Random	42	24.80	13.97	9.02	6.15	42.97	27.05
Random	0	26.64	14.78	9.37	6.34	42.37	26.16
Random	1	24.48	13.81	8.86	6.02	42.87	26.28
Min loss	42	26.79	15.14	9.86	6.82	42.37	27.19
Min loss	0	22.23	12.78	8.35	5.76	**42.98**	27.81
Min loss	1	26.87	15.41	10.03	6.89	42.57	27.44
Evaluation with paraphrases							
No paraphrases	42	33.75	19.78	12.92	8.85	44.02	30.80
No paraphrases	0	32.43	19.20	12.71	8.85	44.46	**31.40**
No paraphrases	1	**34.06**	**19.91**	**12.99**	**8.88**	43.83	30.53
Random	42	30.33	17.75	11.75	7.85	**44.67**	30.08
Random	0	33.03	19.15	12.30	8.30	44.16	29.52
Random	1	29.93	17.49	11.34	7.71	44.62	29.31
Min loss	42	32.51	18.86	12.29	8.43	44.10	29.90
Min loss	0	26.76	15.72	10.90	7.06	44.54	30.29
Min loss	1	32.42	19.09	12.44	8.51	44.21	30.08

Uni-Sign							
Paraphrase mode	Seed	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEURT	Rouge-L
Evaluation without paraphrases							
No paraphrases	42	35.63	22.82	15.89	11.58	45.31	33.12
No paraphrases	0	35.75	22.95	16.03	11.67	45.14	32.92
No paraphrases	1	**36.14**	**23.20**	**16.21**	**11.79**	**45.44**	**33.46**
Random	42	31.98	19.75	13.34	9.39	44.51	29.67
Random	0	31.83	19.74	13.44	9.58	44.57	29.74
Random	1	31.76	19.69	13.33	9.42	44.57	29.94
Min loss	42	35.75	22.67	15.71	11.35	45.22	32.52
Min loss	0	36.13	23.07	16.05	11.66	45.22	32.72
Min loss	1	35.99	22.94	16.04	11.73	44.95	32.53
Evaluation with paraphrases							
No paraphrases	42	41.36	26.56	18.36	13.25	46.62	35.05
No paraphrases	0	41.57	26.74	18.54	13.38	46.50	34.97
No paraphrases	1	41.94	26.99	18.73	13.53	**46.80**	**35.29**
Random	42	38.37	24.18	16.41	11.52	46.09	32.44
Random	0	38.23	24.20	16.53	11.73	46.15	32.53
Random	1	38.28	24.21	16.43	11.53	46.19	32.84
Min loss	42	42.01	26.89	18.58	13.33	46.79	34.79
Min loss	0	**42.32**	**27.22**	18.85	13.56	46.76	34.91
Min loss	1	42.21	27.14	**18.89**	**13.68**	46.46	34.80

Table 4: Performance of our models trained with different seeds and different paraphrase-based training strategies on the How2Sign test set. The checkpoints were evaluated both without and with paraphrases, where metrics are computed for all available paraphrases of each test instance and the highest-scoring paraphrase is selected. `No par.` denotes training with only canonical targets, `Random` indicates random sampling from available paraphrases during training, and `Min loss` selects the paraphrase yielding the minimum training loss per instance. For the `Min loss` configuration, a paraphrased target was selected in approximately 60% of training instances in case of both models.