Target-Side Paraphrase Augmentation for Sign Language Translation with Large Language Models

Pedro Dal Bianco
III-LIDI
Universidad Nacional de La Plata
pdalbianco@lidi.info.unlp.edu.ar

Jean Paul Nunes Reinhold
CDTEC, Federal University of Pelotas
jean.pnr@inf.ufpel.edu.br

Oscar Stanchi
III-LIDI
CONICET
ostanchi@lidi.info.unlp.edu.ar

Facundo Quiroga
III-LIDI
Comision de Investigaciones Cientificas
Universidad Nacional de La Plata
fquiroga@lidi.info.unlp.edu.ar

Franco Ronchetti
III-LIDI
Comision de Investigaciones Cientificas
Universidad Nacional de La Plata
fronchetti@lidi.info.unlp.edu.ar

Ulisses Brisolara Corrêa
Universidade Federal de Pelotas
ub.correa@inf.ufpel.edu.br

Abstract

Sign language translation (SLT) remains constrained by the limited availability of paired sign-video/text corpora and by the heavy-tailed vocabularies typical of real-world datasets. We study a target-side augmentation strategy in which a large language model (LLM) generates controlled paraphrase variants of the reference spoken-language sentence while the sign input remains unchanged. Concretely, we use GPT-4o to produce semantically faithful variants of the training targets and train a Signformer-style pose-based Transformer under a two-stage schedule: pre-training on the augmented corpus followed by fine-tuning on the original references.

We evaluate this strategy on three datasets that span complementary challenges: PHOENIX14T (German Sign Language), a real-world corpus with moderate lexical diversity; the Greek Sign Language Dataset with highly controlled, repetitive recordings; and LSA-T (Argentinian Sign Language), a naturalistic corpus with a large vocabulary and severe long-tail sparsity. This range allows us to characterize precisely when and why target-side augmentation is beneficial.

On PHOENIX14T, augmentation improves BLEU-4 from 9.56 to 10.33, demonstrating that paraphrastic exposure helps the decoder generalize beyond memorized reference phrasing. The near-saturated GSL baseline and the extremely sparse LSA-T setting reveal the limits of the approach: in both cases, single-reference lexical overlap metrics are insufficient to capture the full picture, motivating a complementary semantic evaluation. To our knowledge, this is the first study to examine LLM-generated target-side paraphrases as an augmentation mechanism for SLT, and the first to apply an LLM-as-a-Judge evaluation protocol to SLT. This complementary evaluation reveals gains in semantic fidelity that lexical overlap metrics understate.

1. Introduction

Sign language translation (SLT) aims to map sign-language video directly to spoken-language text, offering a path toward more accessible communication between deaf signers and hearing communities. SLT sits at the intersection of computer vision and natural language processing, and although the field has advanced substantially in recent years [4, 5], robust translation remains difficult.

Signed utterances distribute meaning across manual and non-manual articulators, exhibit strong coarticulation, and vary with signer style and recording conditions. At the same time, large parallel sign-video/text corpora remain scarce [3], and the datasets that do exist often combine narrow domains, modest sample counts, and heavy-tailed target vo-

cabularies. Many words appear only a handful of times, if at all, making it difficult for translation models to learn stable text-generation behavior. The widely used RWTH-PHOENIX-Weather 2014T (PHOENIX14T) benchmark [4], for instance, contains relatively repetitive domain-specific phrasing, whereas broader real-world corpora show much larger vocabularies and many more singletons. The combined effect of multimodal complexity, domain shift, and lexical sparsity remains a central obstacle for SLT.

These constraints also motivate compact, reproducible SLT models. High-capacity video-based systems can deliver stronger absolute performance, but they often depend on costly visual backbones and training pipelines that are difficult to reproduce in smaller research settings. Pose-based models derived from lightweight architectures such as Signformer [19] offer a more practical alternative: by representing each video as a sequence of body and hand landmarks, they reduce input dimensionality and make experimentation more affordable. We adopt that perspective here, using MediaPipe-derived pose features with a Signformer-inspired Transformer so that the effect of data augmentation can be studied within a controlled and resource-efficient setup.

Data augmentation is a standard strategy in low-resource machine translation. In spoken-language MT, techniques such as back-translation and paraphrasing often improve generalization under limited supervision [8, 15]. In sign-language research, however, augmentation has mostly focused on the sign side, either by generating synthetic gloss–text pairs [12] or by synthesizing additional sign videos through sign-language production models [17]. Our focus is different: we augment the *text side* of existing SLT pairs with large language models. More specifically, we study target-side paraphrase augmentation: each reference sentence is rewritten into semantically faithful variants, exposing the decoder to multiple acceptable surface realizations of the same signed content. The goal is to reduce overfitting to rigid reference wording and improve robustness to lexical and syntactic variation at decoding time. To our knowledge, LLM-generated target-side paraphrase augmentation has not previously been studied for SLT, although LLMs have begun to appear in SLT pipelines in other roles, such as stronger text decoders [18].

We evaluate the method on three datasets spanning different sign languages and levels of difficulty: PHOENIX14T (German Sign Language) [4], a weather-forecast corpus with relatively formulaic text; a Greek Sign Language (GSL) educational corpus [1]; and an Argentinian Sign Language corpus derived from LSA-T [2]. Our hypothesis is intentionally narrow: target-side augmentation should help the decoder tolerate superficial lexical and syntactic variation while preserving signed meaning, rather than solve broader challenges such as unseen signers or cross-domain transfer.

This also motivates a semantic evaluation alongside BLEU, which we conduct using an LLM-as-a-Judge protocol described in Section 5.

Our main contributions are as follows: (1) we introduce LLM-based target-side paraphrase augmentation for sign language translation and release three augmented SLT datasets covering DGS, GSL, and LSA; (2) we present a three-dataset study on PHOENIX14T, GSL, and LSA-T, showing that the effect of this augmentation depends strongly on corpus characteristics, ranging from a BLEU-4 gain on PHOENIX14T to negligible or slightly negative effects in simpler or more extreme long-tail settings; and (3) we conduct a semantic evaluation using an LLM-as-a-Judge protocol on PHOENIX14T and GSL, showing that augmentation yields gains in semantic fidelity that lexical overlap metrics understate. All code and datasets are publicly available [1].

2. Related Work

Sign Language Translation. Early SLT systems typically adopted a two-stage pipeline: continuous sign language recognition first predicted an intermediate gloss sequence, which was then translated into spoken text [4]. Glosses provide a convenient intermediate representation, but obtaining them is labor-intensive and they cannot capture the full richness of signed communication, including facial expressions and classifier constructions. More recent work has therefore shifted toward *gloss-free* SLT, learning an end-to-end mapping from video directly to spoken-language text [5, 6]. This setting is more challenging and usually trails gloss-based performance, but it is also more scalable because it requires only video–text pairs. Modern gloss-free systems increasingly rely on transformer architectures and, in some cases, large pretrained models. For example, Sign2GPT [18] combines a pretrained CLIP visual encoder with a GPT-style decoder and lightweight adapters, achieving state-of-the-art results on PHOENIX14T and CSL-Daily. By contrast, [19] proposed *Signformer*, a highly lightweight transformer without pretrained components that still achieved competitive performance. Our model follows that efficiency-oriented line of work, but uses body-pose sequences rather than dense visual embeddings.

Data Augmentation in SLT. The scarcity of sign-to-text data has encouraged a broad range of augmentation methods. Beyond standard video perturbations such as mirroring or spatial jitter, prior SLT work has explored richer sign-side augmentation strategies. One direction synthesizes additional training examples through sign-language production models [16]; another uses motion synthesis, stitching, or generative models such as SignGAN and SignSplat to create new sign video variations, with substantial relative BLEU improvements in some settings [17]. Target-side augmenta-

[1] URL anonymized for review purposes.

tion has received less attention in SLT. [12] expanded gloss-to-text training data through paraphrase pairs derived from monolingual data and heuristic rules. More broadly in NLP, LLMs such as GPT-3 and GPT-4o have been used to generate paraphrases and synthetic examples for low-resource tasks [7]. Our work brings that idea to SLT by rewriting target-language references while keeping the sign input fixed, making it complementary to sign-side augmentation rather than a substitute for it.

3. Methodology

3.1. Model Architecture

Our baseline follows the standard Signformer encoder–decoder Transformer architecture [19] (Figure 1), adapted to accept skeleton keypoints as input in place of visual features. Instead of spatio-temporal visual embeddings extracted from raw video frames, each input video is represented as a sequence of pose keypoints. Specifically, MediaPipe Holistic [11] extracts 2D coordinates for 33 body landmarks, 21 landmarks per hand (left and right), and a facial subset around the mouth and eyebrows. These are concatenated into a per-frame feature vector, producing a time series of pose features that is projected into the model's embedding space through a linear layer before entering the encoder. This representation is motivated by two practical considerations. First, landmark sequences emphasize articulatory motion rather than appearance, suppressing much of the background, lighting, and camera variation that differs across datasets and signers. Second, self-attention is well suited to continuous signing, where semantically relevant cues may be distributed across long temporal windows and multiple channels such as the hands, face, and upper body. In effect, the model can devote more of its capacity to temporal structure and coordinated movement rather than to reconstructing visual texture.

Using skeleton data substantially reduces input dimensionality and removes much of the irrelevant visual clutter, which can make training and inference more practical for resource-constrained environments [19]. The trade-off is that some information is inevitably lost, including fine-grained appearance cues and subtle gestures that are not well captured by keypoints. Prior work suggests that pose-based approaches may therefore lag behind image-based models on unconstrained translation tasks [21]. We treat pose extraction quality as a critical dependency: missing hand tracks, unstable facial landmarks, or brief occlusions can directly degrade the encoder input. In practice, we mitigate these issues with standardized preprocessing and sequence normalization, and we interpret the reported results as comparisons of augmentation policies within a fixed pose-based pipeline. The absence of a heavy visual backbone also simplifies the engineering stack: fewer components must be tuned, memory can be reallocated to longer sequences or larger batches, and preprocessing failures are easier to diagnose at the landmark level. Although our absolute BLEU scores remain below state of the art systems that operate on full video (see Section 5), the relative comparisons with and without augmentation remain informative within this setting.

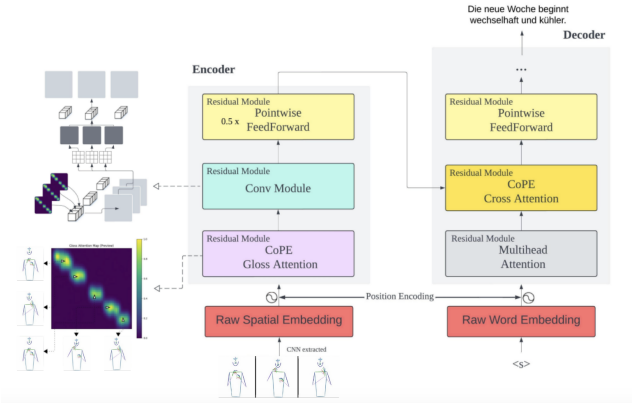

Figure 1. Overview of the adapted *Signformer* architecture (originally from [19]) used in our experiments. Instead of the CNN-extracted frame tokens in the original, each frame's concatenated hand, upper-body, and selected facial landmarks (normalized and linearly projected) feed the encoder directly as pose keypoints.

3.2. LLM-Based Data Augmentation

To augment the training data, we use GPT-4o as a paraphrase generator. For each original video–sentence pair (V, T), where T is the ground-truth spoken-language translation of the sign video V, we generate $N = 3$ additional sentences T'_1, T'_2, T'_3 intended to preserve the meaning of T. The prompt instructs the model to produce semantically faithful paraphrases in the original target language, preserving propositional content while allowing controlled lexical and syntactic variation. This is desirable because the sign input remains fixed while the decoder sees a broader set of valid target realizations.

The prompt requires a structured JSON response, which makes the pipeline easier to validate automatically and easier to resume after interruptions. The instructions explicitly require the model to preserve tense, register, and propositional content, allowing changes in wording or word order only when the meaning remains intact.

For each original sign video V, the resulting training set contains the reference T and three paraphrases T'_1, T'_2, T'_3. During training, these are materialized as four separate examples (V, T), (V, T'_1), (V, T'_2), and (V, T'_3), meaning that the same sign input is paired with multiple textual realizations. Figure 2 summarizes the overall procedure, and Table 1 provides representative examples from PHOENIX14T.

Not every generated sentence is retained. To reduce se-

mantic drift and remove trivial copies, we filter candidate paraphrases with four surface-form similarity measures between the original sentence and each variant: character-level Jaccard overlap, word-level Jaccard overlap, normalized Levenshtein similarity, and trigram overlap. Let $\bar{s}(T, T_i')$ denote the mean of these four scores. We keep a variant only if

$$0.3 \leq \bar{s}(T, T_i') \leq 0.95,$$

and we discard exact duplicates regardless of the threshold. The lower bound rejects variants that are too dissimilar and therefore more likely to have altered the meaning, while the upper bound removes nearly identical rewrites that contribute little diversity. In practice this lightweight filtering step proved important because LLM outputs occasionally oscillated between overly literal copies and overly free reformulations. The generation pipeline also includes checkpointing so that long batch runs can be resumed without regenerating already accepted samples. In production, outputs were required to return structured JSON so that results could be validated automatically before entering the training set; the full prompt template is shown in Figure 2.

3.3. Training Schedule

We compare two conditions:

- **Baseline:** the model is trained only on the original, non-augmented training set.
- **+Augmentation:**
  - *Stage1:* pre-train on the augmented corpus consisting of the original targets plus three GPT-4o paraphrases per instance.
  - *Stage2:* fine-tune on the original training set only, so that the decoder is re-aligned with the reference phrasing and less likely to overproduce rare paraphrastic variants. Unless otherwise stated, all hyperparameters remain identical across conditions, and early stopping is performed on the same development set.

This two-stage schedule is central to the method. Training exclusively on paraphrastic variants can broaden the decoder distribution, but it can also bias the model toward outputs that are semantically valid yet deviate from the single reference used at evaluation time. The final fine-tuning stage partially counteracts that tendency by re-centering the model on the original corpus style without discarding the broader lexical exposure learned during pre-training. For fairness, both conditions use the same optimization policy, validation protocol, and stopping criterion; the intended difference is only the presence or absence of augmented targets. The training setup follows standard sequence-to-sequence practice: teacher forcing with cross-entropy loss, a warm-up-and-decay learning-rate schedule, label smoothing, and early stopping on validation loss.

4. Datasets and Evaluation

We evaluate the proposed approach on three SLT datasets that differ markedly in linguistic diversity, recording conditions, and lexical structure, all of which influence how target-side augmentation behaves.

The **PHOENIX14T** dataset [4] contains weather-broadcast recordings in German Sign Language (DGS) paired with German text. It is a real-world corpus with relatively consistent, domain-specific phrasing and limited topic variation. That repetitiveness simplifies part of the translation problem, but the natural recording conditions still introduce visual variation across signers and sessions, yielding a moderately challenging benchmark.

The **GSL** dataset [1], in contrast, is recorded under controlled laboratory conditions and features a small set of signers repeatedly producing a restricted inventory of predefined sentences. As a result, it has low linguistic and visual variability, high redundancy across samples, and almost no rare target tokens. Models can therefore memorize sentence patterns easily and achieve near-perfect BLEU, but this simplicity also makes the benchmark sensitive to any increase in output variation.

Finally, **LSA-T** [2] consists of real-world videos collected from diverse sources, with substantial variation in signers, lighting, and signing style. Its naturalistic content and broad Spanish vocabulary make it much harder than the previous two datasets. The high proportion of singleton words and irregular phrasing creates a pronounced long-tail distribution, leading to sparse lexical coverage and very low baseline translation accuracy. This makes LSA-T a useful test bed for assessing whether augmentation can help under severe data sparsity.

Table 2 quantifies the key differences across these three settings.

For all datasets, videos are preprocessed with MediaPipe to extract pose sequences, as described above. Coordinate values are normalized per frame sequence and frame rates are standardized for model input following steps similar to [20], while the textual side is lowercased and tokenized consistently across corpora. Translation quality is measured with case-insensitive BLEU-4 [13] on the official test splits.

5. Results and Analysis

5.1. Main Three-Dataset Benchmark

Table 3 reports BLEU-4 on the test sets for the baseline model and the two-stage **+Augmentation** configuration.

Overall trends. PHOENIX14T shows an improvement of +0.77 BLEU. In a moderately rich yet still formulaic domain, exposure to paraphrastic reorderings appears to help the decoder generalize beyond memorized templates, while

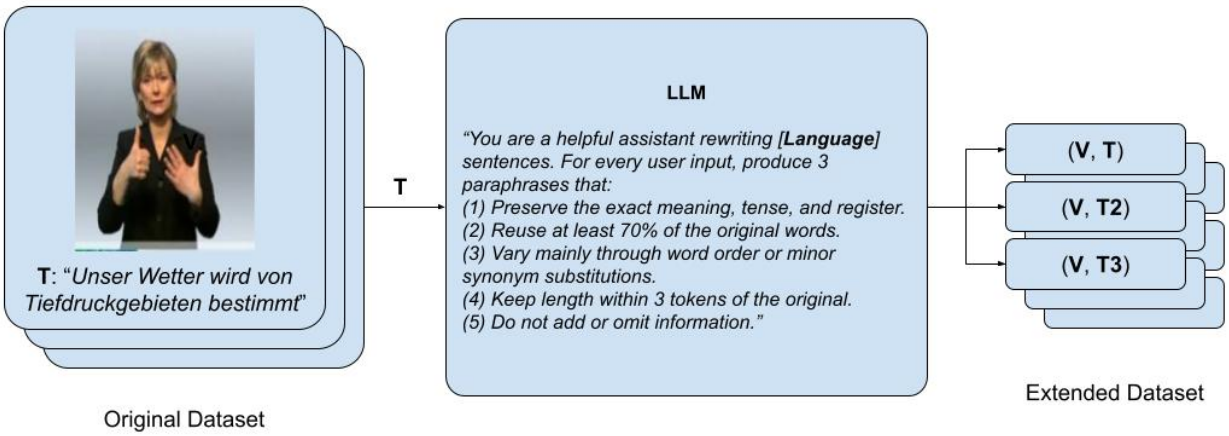

Figure 2. Overview of the LLM-augmented SLT pipeline. For each video–text pair (V, T), an LLM generates three paraphrases (T'_1, T'_2, T'_3) that preserve meaning while introducing controlled lexical and syntactic variation. Training then follows two stages: pre-training on the augmented corpus and fine-tuning on the original targets only. At inference time, the model translates directly from sign input to text.

Original (reference)	**LLM paraphrases**
tiefdruckgebiete bestimmen unser wetter *low-pressure areas determine our weather*	• Unser Wetter wird von Tiefdruckgebieten bestimmt. *Our weather is determined by low-pressure areas.* • Die Bestimmung unseres Wetters erfolgt durch Tiefdruckgebiete. *The determination of our weather is due to low-pressure areas.*
auch mit den temperaturen geht es aufwärts *the temperatures are also rising*	• Auch die Temperaturen steigen an. *The temperatures are also increasing.* • Die Temperaturen gehen ebenfalls nach oben. *The temperatures are also going up.*
eine gewitterfront überquert deutschland von west nach ost *a thunderstorm front crosses Germany from west to east*	• Eine Gewitterfront zieht von Westen nach Osten über Deutschland. *A thunderstorm front moves from west to east across Germany.* • Von Westen nach Osten überquert eine Gewitterfront Deutschland. *From west to east, a thunderstorm front crosses Germany.*

Table 1. Examples of GPT-4o paraphrases paired with original PHOENIX14T training references.

the final fine-tuning stage helps keep the output close to the evaluation style. By contrast, the GSL benchmark starts from an exceptionally high baseline of 94.38 BLEU, reflecting substantial overlap and low linguistic variability between training and test. In that near-saturated setting, augmentation slightly reduces performance to 92.22 BLEU: the decoder learns alternative phrasings that remain semantically valid but do not exactly match the single reference, and fine-tuning does not completely suppress those variants. Finally, the LSA-T setting remains extremely low in both conditions, at roughly 1.2 BLEU. Because the paraphrasing prompt deliberately preserves the same rare content words, target-side augmentation does not address the underlying problem of severe data sparsity on the sign side.

Data characteristics matter. The utility of LLM paraphrasing is closely tied to vocabulary breadth and to the prevalence of infrequent tokens. When a dataset provides enough lexical variety, as in PHOENIX14T, paraphrastic exposure can help the model tolerate alternative word orderings and light lexical substitutions at test time. When the task is unusually simple, as in the GSL benchmark, greater output diversity can reduce single-reference BLEU even if the underlying meaning is preserved. When the vocabulary is extremely sparse, as in the LSA-T setting, paraphrasing the target text alone does not solve the central coverage problem: many content words and corresponding sign patterns remain too rare for the decoder to learn robustly.

Why BLEU can understate gains. Target-side augmentation creates a methodological mismatch at evaluation time. During training, the model is encouraged to treat multiple textual realizations of the same signed content as acceptable. BLEU, however, rewards overlap with only one reference sentence. For a weather forecast, training may expose the decoder to alternatives such as "morgen wird es regnen," "es

Statistic	PHOENIX14T (DGS)	GSL	LSA-T (LSA)
Language (target)	German	Greek	Spanish
Sign language	DGS	GSL	LSA
Real-world footage	Yes	No	Yes
No. of signers	9	7	103
Duration [h]	10.71	9.51	21.78
Samples (clips)	7,096	10,295	8,459
Unique sentences	5,672	331	8,102
% unique sentences	79.93%	3.21%	95.79%
Vocabulary size (types)	2,887	N/A	14,239
Singletons (types with count=1)	1,077	0	7,150
% singletons	37.3%	0%	50.21%
Resolution	210×260	848×480	1920×1080
FPS	25	30	30

Table 2. Corpus statistics for the three datasets used in the main experiments. The lower block highlights lexical properties related to long-tail behavior.

Dataset	Baseline (BLEU-4)	+Augmentation (BLEU-4)
PHOENIX14T (DGS)	9.56	10.33
GSL (Greek)	94.38	92.22
LSA (Spanish)	1.18	1.19

Table 3. Test BLEU-4 for baseline training and LLM-augmented training on the three evaluation datasets.

regnet morgen," and "für morgen ist Regen vorhergesagt." If the model later produces another semantically correct variant that differs lexically from the single test reference, BLEU will penalize it despite preserving the meaning. The objective encouraged by augmentation is therefore broader than the one measured by single-reference lexical overlap, which helps explain why a modest BLEU gain can still correspond to a more meaningful improvement in semantic robustness.

5.2. Semantic Evaluation via LLM-as-a-Judge

The lexical-overlap mismatch discussed above motivates a semantic evaluation of the PHOENIX14T and GSL runs from the previous section. An LLM judge scores translations along semantic fidelity and linguistic quality, counts *fluent but wrong* outputs, and performs direct pairwise comparisons between baseline and augmented hypotheses.

Recent literature supports this direction. Strong LLM judges have shown high agreement with human preferences on open-ended generation benchmarks [22], and LLM-based translation evaluators have achieved competitive or state-of-the-art correlation with human judgments in machine translation [9]. More structured prompting strategies such as G-Eval [10] further suggest that dimension-specific scoring can make automated judging more interpretable. At a lighter-weight end of the spectrum, embedding-based methods such as Sentence-BERT [14] provide an additional semantic signal between pure lexical overlap and full LLM judging.

Dataset	Baseline	Augmented	Change
PHOENIX14T	2.51	3.65	+45.0%
GSL	7.72	8.77	+13.6%

Table 4. LLM-as-a-Judge semantic fidelity scores (GPT-5.2) for baseline and augmented models.

Results. Table 4 reports semantic fidelity on both datasets. On PHOENIX14T, augmentation raises fidelity from 2.51 to 3.65 (+45%). Additional PHOENIX14T metrics reinforce this: the rate of fluent-but-semantically-incorrect translations drops from 54.8% to 35.5% (-19.3 pp), and pairwise preference judgments favor the augmented model in 52.9% of comparisons versus 13.1% for the baseline. On GSL, where the baseline was already strong, fidelity still improves from 7.72 to 8.77 (+13.6%), confirming semantic gains in a regime where lexical overlap metrics are near-saturated. Together, these results confirm that target-side augmentation improves semantic quality beyond what BLEU alone captures.

Limitations of semantic judging. Any LLM-based evaluator must be used cautiously. In our experiments, GPT-5.2 served as the judge while GPT-4o generated the augmented training targets. These are architecturally distinct (GPT-5.2 is a reasoning model), which reduces but does not eliminate this concern: shared training lineage may still introduce subtle stylistic alignment between generator and evaluator. For that reason, semantic judging should be treated as complementary evidence and ideally validated with human assess-

ment or at least triangulated with simpler automatic signals. This concern is particularly relevant in SLT, where small lexical differences can be acceptable while subtle semantic errors remain difficult to detect automatically.

6. Conclusion

Target-side LLM augmentation is not uniformly beneficial for SLT: its value depends on where a corpus sits along the spectrum from formulaic to long-tail sparse. On PHOENIX14T, paraphrastic exposure improves BLEU and reduces semantically incorrect fluent outputs. On GSL, the near-perfect baseline leaves no room for lexical variation to help and single-reference BLEU penalizes it. On LSA-T, target-side rewriting cannot address sign-side sparsity. The central lesson is that augmentation works when a dataset has enough lexical breadth for variation to improve generalization, but not so little diversity that alternative valid phrasings are penalized at evaluation time.

The semantic evaluation confirms that gains from augmentation exceed what BLEU alone captures, while the architectural gap between GPT-4o (augmentation) and GPT-5.2 (judge, a reasoning model) reduces but does not eliminate the risk of shared stylistic bias; future work should validate with human judgments or cross-family evaluators.

Promising directions include combining target-text augmentation with sign-side augmentation so that linguistic and motion variation are learned jointly; evaluating with multiple references or human judgments, where semantically valid paraphrases are less likely to be penalized; and replacing GPT-4o with open-source LLMs or task-specific paraphrasers to clarify whether the gains stem from the augmentation principle or from the specific generator.

References

[1] Nikolas Adaloglou, Theocharis Chatzis, Ilias Papastratis, Andreas Stergioulas, Georgios Th Papadopoulos, Vassia Zacharopoulou, George J Xydopoulos, Klimnis Atzakas, Dimitris Papazachariou, and Petros Daras. A comprehensive study on sign language recognition methods. *arXiv preprint arXiv:2007.12530*, 2020.

[2] Pedro Dal Bianco, Gast'on R'ıos, Franco Ronchetti, Facundo Quiroga, Oscar Stanchi, Waldo Hasperu'e, and Alejandro Rosete. Lsa-t: The first continuous argentinian sign language dataset for sign language translation. In *Advances in Artificial Intelligence – IBERAMIA 2022*, page 293–304. Springer, Cham, 2022.

[3] Danielle Bragg, Oscar Koller, Miriam Bellard, Larwan Berke, Naomi Caselli, and etc. Sign language recognition, generation, and translation: An interdisciplinary perspective. *ACM Transactions on Accessible Computing*, 12(2):5:1–5:44, 2019.

[4] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7784–7793, 2018.

[5] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 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 10020–10030, 2020.

[6] Shizhe Chen, Yuecong Wang, and etc. Two stream transformer networks for sign language translation. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[7] Ehsan Davoodi et al. Improving low-resource classification via large language models for data augmentation. In *Proceedings of the 60th Annual Meeting of the ACL (Short Papers)*, 2022.

[8] Xinyu Hu et al. Text data augmentation made simple by leveraging llms: A case study on low-resource nlu tasks. In *Proceedings of the EMNLP 2021 (Findings)*, 2021.

[9] Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland, 2023. European Association for Machine Translation.

[10] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, 2023. Association for Computational Linguistics.

[11] Wesley Maia, António M. Lopes, and Sérgio A. David. Automatic sign language to text translation using mediapipe and transformer architectures. *Neurocomputing*, 642:130421, 2025.

[12] Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. Data augmentation for sign language gloss translation. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL) at MTSummit*, pages 1–11, Virtual, 2021. Association for Machine Translation in the Americas.

[13] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[14] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics.

[15] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the ACL*, 2016.

[16] Stefanie Stoll et al. Text2sign: Towards sign language production using neural machine translation and generative adversar-

ial networks. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[17] Harry Walsh, Maksym Ivashechkin, and Richard Bowden. Using sign language production as data augmentation to enhance sign language translation. *arXiv preprint arXiv:2506.09643*, 2025.

[18] Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. Sign2gpt: Leveraging large language models for gloss-free sign language translation. In *International Conference on Learning Representations (ICLR)*, 2024.

[19] Eta Yang. Signformer is all you need: Towards edge ai for sign language. *arXiv preprint arXiv:2411.12901*, 2024.

[20] Tomáš Železný, Jakub Straka, Václav Javorek, Ondřej Valach, Marek Hrúz, and Ivan Gruber. Exploring pose-based sign language translation: Ablation studies and attention insights. *arXiv preprint arXiv:2507.01532*, 2023.

[21] Tomáš Zelezný, Jakub Straka, Václav Javorek, Ondřej Valach, Marek Hrúz, and Ivan Gruber. Exploring pose-based sign language translation: Ablation studies and attention insights. *arXiv preprint arXiv:2507.01532*, 2025.

[22] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.