Finetuning Pre-trained Language Models for Bidirectional Sign Language Gloss to Text Translation

Arshia Kermani, Habib Irani, Vangelis Metsis
Department of Computer Science
Texas State University
San Marcos, TX 78666, USA
{arshia.kermani, habibirani, vmetsis}@txstate.edu

Abstract

Sign Language Translation (SLT) is a crucial technology for fostering communication accessibility for the Deaf and Hard-of-Hearing (DHH) community. A dominant approach in SLT involves a two-stage pipeline: first, transcribing video to sign language glosses, and then translating these glosses into natural text. This second stage, gloss-to-text translation, is a challenging, low-resource machine translation task due to data scarcity and significant syntactic divergence. While prior work has often relied on training translation models from scratch, we show that fine-tuning large, pre-trained language models (PLMs) offers a more effective and data-efficient paradigm. In this work, we conduct a comprehensive bidirectional evaluation of several PLMs (T5, Flan-T5, mBART, and Llama) on this task. We use a collection of popular SLT datasets (RWTH-PHOENIX-14T, SIGNUM, and ASLG-PC12) and evaluate performance using standard machine translation metrics. Our results show that fine-tuned PLMs consistently and significantly outperform Transformer models trained from scratch, establishing new state-of-the-art results. Crucially, our bidirectional analysis reveals a significant performance gap, with Text-to-Gloss translation posing a greater challenge than Gloss-to-Text. We conclude that leveraging the linguistic knowledge of pre-trained models is a superior strategy for gloss translation and provides a more practical foundation for building robust, real-world SLT systems.

1 Introduction

Automatic Sign Language Translation (SLT) is a vital research field focused on bridging communication barriers for the millions of individuals in the Deaf and Hard-of-Hearing (DHH) community (Bragg et al., 2019). The development of robust SLT systems has profound implications for social inclusion, education, and access to essential services, particularly in domains like telehealth where the availability of human interpreters can be limited (Pikoulis et al., 2022).

A dominant paradigm in SLT research decomposes the complex video-to-text translation problem into a more manageable two-stage pipeline (Camgoz et al., 2018). First, a Sign Language Recognition (SLR) module analyzes the input video to generate a sequence of textual labels, known as "glosses." These glosses represent the individual signs in their original signed order. Second, a machine translation module translates this sequence of glosses into a grammatically correct natural language sentence. This paper focuses on this critical second stage: the bidirectional translation between sign language glosses and natural language text (Gloss $\Leftrightarrow$ Text).

The task of translating sign glosses, however, presents unique challenges for Neural Machine Translation (NMT). Glosses are an intermediate representation that simplifies the visual signal into a text-like sequence, but they omit many linguistic features and non-manual markers (e.g., facial expressions). While the lexicon of glosses often overlaps significantly with the target natural language, their syntax follows the grammatical rules of the source sign language, which can be vastly different. For example, American Sign Language (ASL) has a distinct word order and grammatical structure from English (Sandler and Lillo-Martin, 2006). This results in a translation task characterized by high lexical overlap but significant syntactic divergence. Compounding this challenge, the parallel gloss-text corpora available for training are typically small, making this an extremely low-resource NMT problem (Yin and Read, 2020).

Previous neural approaches have demonstrated the viability of the Transformer architecture for this task, but have primarily relied on training models from scratch on these limited datasets (Yin and Read, 2020). We hypothesize that this approach is data-inefficient and that a more effective strategy is

Proceedings of the Workshop on Sign Language Processing (WSLP), pages 77–85

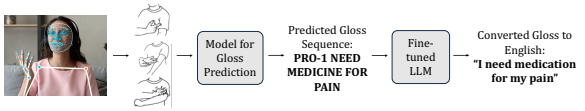

Figure 1: The two-stage Sign Language Translation (SLT) pipeline. This work focuses on the second stage: translating sign language gloss sequences into natural language text and vice-versa. The example shows ASL glosses being translated into an English sentence.

to leverage the vast linguistic knowledge encoded in large, pre-trained language models (PLMs).

Recently, the focus has begun to shift towards fine-tuning LLMs, with work such as (Fayyazsanavi et al., 2024) achieving strong results by developing specialized techniques like novel loss functions and data augmentation for the unidirectional Gloss-to-Text task. Our work complements these efforts by asking a different, foundational question: how do various modern PLMs and architectures perform across the full, bidirectional translation pipeline? By fine-tuning these models, which have already learned the rich grammatical and semantic nuances of the target language from massive text corpora, we can adapt them to the specific task of gloss translation more effectively.

The main contributions of this work are as follows:

- We conduct the first large-scale, systematic comparison of fine-tuning various modern PLMs, including T5, Flan-T5, mBART, and Llama, for the bidirectional gloss-to-text and text-to-gloss translation tasks.
- We empirically demonstrate that our fine-tuning approach significantly outperforms the strong baseline of a Transformer trained from scratch, establishing new state-of-the-art results on the RWTH-PHOENIX-14T, SIGNUM, and ASLG-PC12 benchmarks.
- We provide a comparative analysis of different model architectures (encoder-decoder vs. decoder-only) and pre-training paradigms (e.g., instruction-tuning) to identify the most suitable approaches for this unique translation task.
- We will release our fine-tuned models and experimental code to the research community to foster reproducibility and accelerate future progress in SLT.

2 Related Work

Language Models are increasingly applied across diverse domains, including label quality improvement (Mahjourian and Nguyen, 2025), Sentiment Analysis (Mohammadagha et al., 2025), secure software development practices (Torkamani et al., 2025), and mental health text analysis (Kermani et al., 2025). They have also shown growing potential in advancing translation tasks such as SLT.

2.1 Sign Language Gloss-to-Text Translation

The translation of sign language glosses to natural language text has been an active area of research within SLT. Early approaches often relied on rule-based systems or statistical machine translation (SMT) methods. For instance, the widely-used ASLG-PC12 dataset was itself generated using a rule-based, part-of-speech-based grammar to convert English text into ASL glosses (Othman and Jemni, 2012). However, these methods often struggle to capture the fluency and complexity of natural language.

With the advent of deep learning, the focus shifted to neural machine translation (NMT) models. An initial line of work applied Recurrent Neural Network (RNN) based architectures with attention to the task (Camgoz et al., 2018). A significant step forward was made by (Yin and Read, 2020), who demonstrated the effectiveness of the Transformer architecture (Vaswani et al., 2017) for this task. Their work, which serves as a primary baseline for our study, involved training Transformer models *from scratch* on gloss-text corpora like RWTH-PHOENIX-14T and ASLG-PC12. They showed that this approach could achieve state-of-the-art results, establishing a strong benchmark for neural-based gloss-to-text translation.

The inherent low-resource nature of the problem has also inspired other lines of research, such as data augmentation. For example, (Moryossef et al., 2021) proposed rule-based heuristics to generate pseudo-parallel gloss-text pairs from monolingual text to augment the limited training data. While effective, our work explores a complementary direction: instead of augmenting the data, we propose using more powerful models that are better equipped to learn from sparse data.

Concurrent to our work, (Fayyazsanavi et al., 2024) also explore fine-tuning LLMs for Gloss-to-Text translation. Their primary contributions are the development of tailored data augmenta-

tion techniques (paraphrasing and back-translation) and a novel Semantically Aware Label Smoothing (SALS) loss function to handle gloss ambiguities. Their work demonstrates significant improvements on the PHOENIX-2014T dataset. Our research differs in three key aspects: (1) Scope: We conduct a bidirectional analysis, evaluating both Gloss-to-Text (G2T) and Text-to-Gloss (T2G) tasks, whereas their work focuses solely on G2T. (2) Contribution Type: Our work provides a broad, systematic comparison of multiple PLM families and architectures to establish foundational benchmarks, while their work focuses on developing novel, task-specific techniques for a single model. (3) Evaluation Breadth: We validate our findings across three distinct datasets (RWTH-PHOENIX-14T, SIGNUM, and ASLG-PC12) to ensure generalizability, whereas their experiments are conducted on the PHOENIX-2014T dataset.

2.2 Pre-trained Language Models for NMT

The dominant paradigm in modern Natural Language Processing (NLP) has shifted from training task-specific models from scratch to a pre-train and fine-tune approach (Devlin et al., 2019). Large-scale language models like T5 (Raffel et al., 2020), BART (Lewis et al., 2020), and the Llama family (Aaron Grattafiori, 2024) are first pre-trained on vast, web-scale corpora of unlabeled text. During this phase, they learn rich, general-purpose representations of syntax, semantics, and world knowledge.

This pre-trained knowledge can then be transferred to downstream tasks via a second, much shorter fine-tuning phase on a smaller, labeled dataset. This paradigm has proven exceptionally effective for low-resource NMT (Zoph et al., 2016). Instead of learning the target language's grammar and semantics from a small parallel corpus, the model only needs to learn the *mapping* between the source and target representations. Our work is the first to systematically apply and evaluate this powerful paradigm across a diverse set of modern PLMs for the unique challenges of bidirectional sign language gloss translation.

3 Experimental Setup

We designed a comprehensive experimental setup to rigorously evaluate the performance of fine-tuned pre-trained language models (PLMs) against a from-scratch baseline on bidirectional gloss-text translation. Our setup is standardized across all models to ensure fair and reproducible comparisons. The complete code implementation and benchmarks are made publicly available for reproducibility: `anonymized`.

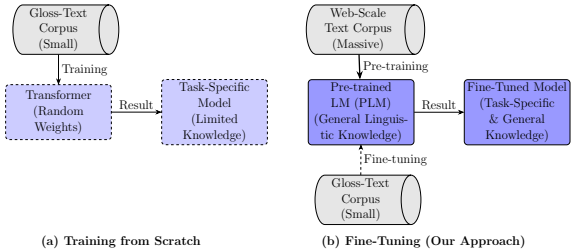

Figure 2: Conceptual comparison of the two training paradigms. (a) The baseline approach trains a Transformer from scratch using only the limited gloss-text corpus. (b) Our approach leverages a large language model pre-trained on vast text corpora and then fine-tunes it on the gloss-text corpus.

3.1 Task Definition

We address two primary translation tasks in this work, treating both as sequence-to-sequence problems:

1. **Gloss-to-Text (G2T):** The model takes a sequence of sign language glosses as input (e.g., `PRO-1 NEED MEDICINE PAIN`) and must generate a grammatically correct sentence in the target natural language (e.g., "I need medicine for the pain.").

2. **Text-to-Gloss (T2G):** The model takes a natural language sentence as input and must generate the corresponding sequence of glosses, reflecting the word order and lexical choices of the target sign language.

3.2 Datasets

We conduct experiments on three publicly available corpora, each with unique characteristics that test different aspects of our models. A summary of the datasets after standard train/dev/test splitting is provided in Table 1.

Dataset	Language Pair	Domain	Train/Dev/Test
PHOENIX	DGS / German	Weather	7,096 / 518 / 642
SIGNUM	DGS / German	Varied	603 / 177 / —
ASLG-PC12	ASL / English	Synthetic	500k / 5k / 5k

Table 1: Overview of datasets. DGS stands for German Sign Language; ASL for American Sign Language. The SIGNUM test set is used for validation.

- **RWTH-PHOENIX-Weather 2014T (Phoenix14T)** (Camgoz et al., 2018) is a widely-used benchmark for continuous sign language research, consisting of German weather forecasts and their corresponding German Sign Language (DGS) gloss transcriptions.
- **SIGNUM** (von Agris and Kraiss, 2010) is a smaller DGS corpus with a more controlled vocabulary, providing a different data condition. We use the original train-test split in our evaluation.
- **ASLG-PC12** (Othman and Jemni, 2012) is a large-scale, synthetically generated corpus of English sentences from Project Gutenberg automatically converted into ASL glosses. While synthetic, its size allows for testing model scalability. We use a 500k-pair subset for training.

3.3 Models and Implementation

We evaluate a from-scratch baseline against four different PLMs.

- **Transformer Baseline (65M params):** For comparison against pre-trained language models (PLMs), we implemented a custom Transformer architecture trained from scratch on the sign language gloss translation tasks.

  The model uses a 4-layer encoder and 4-layer decoder, each with $d_{\text{model}} = 256$ hidden units, 8 attention heads, and a feed-forward dimension of 1024. Positional encodings are added to the token embeddings, and residual connections with dropout (0.2) are applied throughout. To improve parameter efficiency, the output projection layer shares weights with the target embeddings.
- **T5-base (220M params):** A versatile encoder-decoder PLM pre-trained on a text-to-text objective (Raffel et al., 2020).
- **Flan-T5-base (220M params):** An instruction-tuned version of T5, which has been shown to improve zero-shot and few-shot performance on unseen tasks.
- **mBART 50 (610M params):** A multilingual sequence-to-sequence model pre-trained with a denoising objective, which may be particularly suited to handling the ungrammatical nature of glosses (Lewis et al., 2020).
- **Llama 3 8B:** A powerful, modern, decoder-only LLM used to assess the performance of this architectural class (Aaron Grattafiori, 2024).

All models were trained using the HuggingFace Transformers library . For fine-tuning the PLMs, we used the AdamW optimizer with a learning rate of 3×10^{-4} and a batch size of 32. We employed a linear learning rate scheduler with 100 warmup steps and trained for a maximum of 10 epochs with early stopping based on validation loss. For encoder-decoder models, input sequences were prefixed with a task description, e.g., "translate Gloss to English: [GLOSS SEQUENCE]".

3.4 Evaluation Metrics

To provide a comprehensive assessment of translation quality, we use a suite of standard automatic metrics:

- **BLEU** (Papineni et al., 2002): Measures n-gram precision, a standard metric for machine translation quality.
- **ROUGE-L** (Lin, 2004): Measures the longest common subsequence, capturing recall-oriented aspects of the translation.
- **METEOR** (Banerjee and Lavie, 2005): An alignment-based metric that considers synonymy and stemming for a more semantically-aware evaluation.
- **Word Error Rate (WER):** Measures the number of substitutions, deletions, and insertions required to transform the hypothesis into the reference. It is particularly useful for the T2G task where output structure is more rigid.

All scores are computed using the SacreBLEU library (Post, 2018) to ensure consistent and reproducible results. Each experiment was run 10 times with different random initializations.

4 Results and Analysis

We present the results of our experiments on both the Gloss-to-Text (G2T) and Text-to-Gloss (T2G) translation tasks. Our analysis focuses on comparing the performance of fine-tuned pre-trained models against the from-scratch Transformer baseline.

Dataset	Model	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑	ROUGE-L ↑	METEOR ↑	WER ↓
RWTH-PHOENIX-14T	Transformer (Baseline)	34.01	23.71	17.24	13.06	34.75	29.51	81.44
	T5-base	48.71	37.16	30.13	22.73	35.04	31.32	34.37
	Flan-T5-base	45.94	32.68	25.29	19.03	33.33	30.06	36.68
	mBART	58.16	45.86	36.52	25.58	46.30	42.26	26.56
	Llama 8B	**63.56**	**53.45**	**43.78**	**29.92**	**53.33**	**49.14**	**21.32**
SIGNUM	Transformer (Baseline)	59.60	47.26	39.76	34.24	61.22	53.09	46.45
	T5-base	71.21	66.09	60.70	52.87	**86.34**	71.64	22.09
	Flan-T5-base	68.12	64.84	59.45	50.72	85.95	73.83	18.45
	mBART	**82.81**	**77.07**	**72.38**	**67.60**	84.80	**79.68**	**17.61**
	Llama 8B	80.56	75.89	70.35	65.78	82.24	78.92	18.23
ASLG-PC12	Transformer (Baseline)	79.28	73.13	67.75	62.81	89.40	80.60	23.41
	T5-base	91.02	81.90	74.82	68.69	89.17	85.63	20.92
	Flan-T5-base	86.38	74.11	64.91	65.40	84.76	82.64	26.81
	mBART	94.55	90.27	86.08	79.58	92.99	88.03	19.31
	Llama 8B	**96.06**	**91.55**	**87.06**	**83.10**	**94.12**	**90.24**	**17.83**

Table 2: Gloss-to-Text (G2T) translation results on RWTH-PHOENIX-14T, SIGNUM, and ASLG-PC12.

Dataset	Model	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑	ROUGE-L ↑	METEOR ↑	WER ↓
RWTH-PHOENIX-14T	Transformer (Baseline)	33.47	19.06	11.26	6.98	36.72	25.21	82.46
	T5	**60.30**	27.19	15.17	8.49	49.18	40.85	66.90
	Flan-T5	47.50	25.60	15.00	10.00	45.20	38.50	69.30
	mBART	50.10	30.45	19.20	12.10	44.32	36.51	64.10
	Llama	58.43	**38.32**	**25.54**	**16.81**	**51.25**	**44.63**	**61.45**
SIGNUM	Transformer (Baseline)	61.73	49.65	43.26	**37.51**	64.50	55.31	43.93
	T5	72.15	**56.82**	**43.64**	34.66	68.50	58.90	38.84
	Flan-T5	69.23	54.32	41.50	32.44	65.83	56.22	41.21
	mBART	**75.42**	49.54	37.07	25.43	**70.20**	**61.30**	**37.83**
	Llama	62.53	47.61	35.43	29.74	68.63	56.25	45.72
ASLG-PC12	Transformer (Baseline)	82.21	75.47	68.12	64.12	89.77	81.93	23.10
	T5-base	64.20	44.76	31.34	21.73	76.21	60.66	42.13
	Flan-T5-base	43.41	30.87	23.47	18.51	60.58	52.55	55.68
	mBART	73.76	53.41	38.49	27.68	80.65	67.24	35.80
	Llama	**85.64**	**76.78**	**70.62**	**66.33**	**89.91**	**83.85**	**22.37**

Table 3: Text-to-Gloss (T2G) translation results on RWTH-PHOENIX-14T, SIGNUM, and ASLG-PC12.

4.1 Gloss-to-Text Performance

Table 2 summarizes the performance of all models on the G2T task across the datasets. All scores are averaged over 10 runs. Best scores per metric are in **bold**. The results provide strong evidence for our primary hypothesis.

The results of the gloss-to-text experiments across RWTH-PHOENIX-14T, SIGNUM, and ASLG-PC12 shown in Table 2 demonstrate consistent improvements of pre-trained models over the baseline Transformer trained from scratch. On PHOENIX-14T, all PLMs achieve substantial gains in BLEU-4, with the model mBART reaching 25.58 compared to 13.06 for the baseline. The larger Llama 8B extends this advantage further with 29.92 BLEU-4, underscoring the benefit of large-scale pre-training even in low-resource conditions. On SIGNUM, mBART attains 67.60 BLEU-4, while Llama 8B maintains competitive results. For ASLG-PC12, where the dataset is larger and synthetic, Llama 8B achieves the highest score with 83.10 BLEU-4, indicating that decoder-only models are able to fully exploit large-scale parallel data. Overall, the results confirm that fine-tuning PLMs yields not only higher accuracy but also more fluent and grammatically complete translations across diverse data conditions.

4.2 Text-to-Gloss Performance

For the reverse task of Text-to-Gloss (T2G), we evaluate the models' ability to generate syntactically correct gloss sequences. The results are presented in Table 3.

Compared to gloss-to-text, BLEU scores are generally lower and WER is higher, reflecting the structural difficulty of generating gloss sequences that require word deletion, reordering, and strict adherence to gloss grammar. On PHOENIX-14T, the best-performing model achieves only 16.81 BLEU-4, showing the sharp contrast with gloss-to-text performance. On SIGNUM, pre-trained models again outperform the baseline, with T5 and mBART reaching mid-30 BLEU-4 scores, but still below their gloss-to-text counterparts. On ASLG-PC12, Llama achieves the strongest performance with 66.33 BLEU-4, benefiting from the scale of training data, though this remains substantially

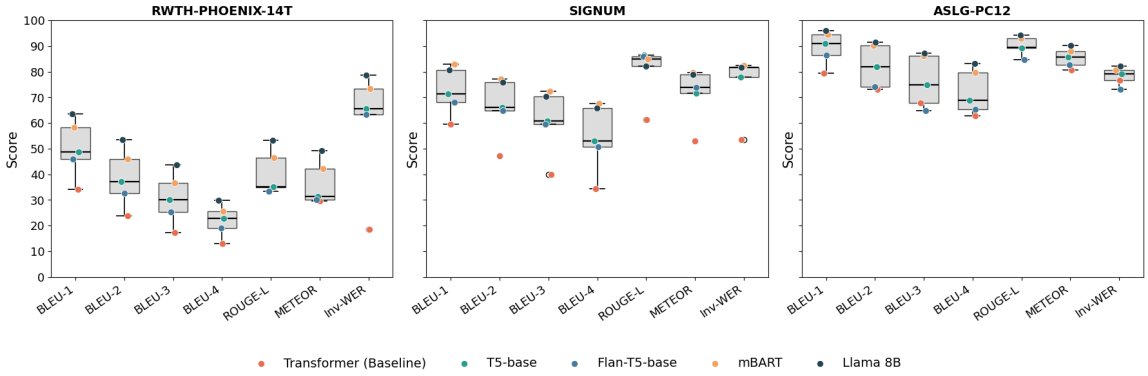

Figure 3: G2T multi-metric comparison across datasets (higher is better; WER inverted).

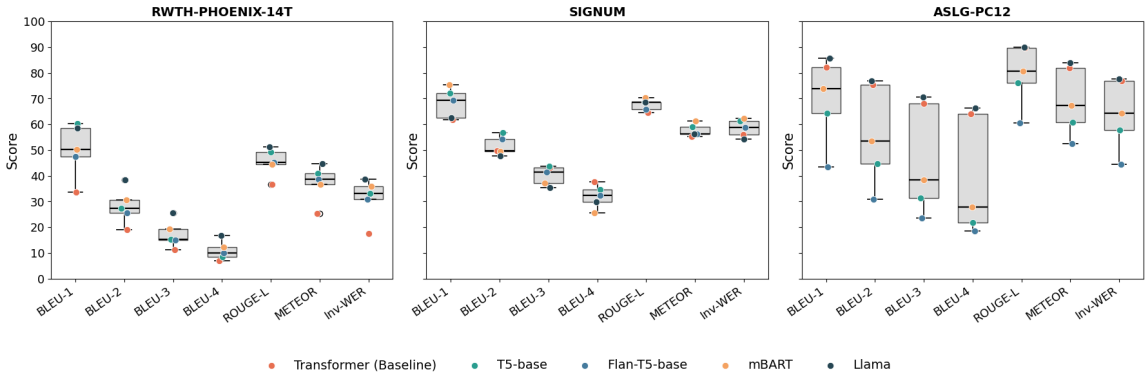

Figure 4: T2G multi-metric comparison across datasets (higher is better; WER inverted).

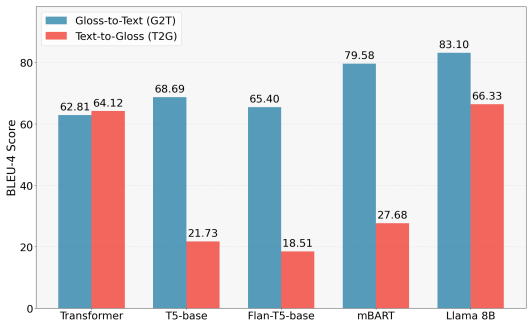

Figure 5: Bidirectional Performance Asymmetry (BLEU-4) on the ASLG-PC12 Dataset.

lower than its gloss-to-text result. These findings confirm the expected asymmetry between the two directions and indicate that text-to-gloss translation is likely to remain a bottleneck in practical bidirectional systems unless further task-specific modeling innovations are introduced. As shown in Figure 5, we observe a clear bidirectional performance asymmetry on the ASLG-PC12 dataset. While G2T achieves higher BLEU, ROUGE-L, and METEOR scores, the corresponding T2G direction results in substantially lower scores across all metrics, highlighting the inherent difficulty of the reverse translation task. A notable exception arises in T2G. As shown in Table3, the from-scratch Transformer achieves the strongest result on SIGNUM and is second only to Llama 8B on ASLG-PC12, outperforming smaller PLMs. This pattern suggests that pre-training for fluent text can conflict with generating non-fluent, rule-like gloss targets.

4.3 Qualitative Analysis

To provide a more intuitive understanding of the performance gap, Table 4 shows example translations from our Flan-T5-base versus the baseline Transformer.

The examples clearly illustrate the advantage of pre-training. The baseline model often produces grammatically incomplete or "telegraphic" text, closely mirroring the structure of the input glosses. In contrast, the fine-tuned Flan-T5 model successfully infers the correct grammatical structure, inserting necessary function words (e.g., "will be", "there is"), handling verb tenses, and producing overall more natural and fluent sentences. This qualitative difference highlights that pre-trained models do not just learn a word-for-word mapping but leverage their internal linguistic models to perform true translation.

Source Text: we are talking about children , the most precious resource that we should protect.
Reference Gloss: X-WE BE TALK ABOUT CHILD , MOST DESC-PRECIOUS RESOURCE THAT X-WE SHOULD PROTECT. **Predicted Gloss:** X-WE BE TALK ABOUT CHILD , MOST FINISH RESOURCE THAT X-WE SHOULD PROTECT.
Source Gloss: IX-1P NOT-YET SEE MOVIE BUT FRIEND RECOMMEND
Reference: I haven't seen the movie yet, but my friend recommended it. **Baseline Output:** I not see movie but friend say good. **Our Output (Flan-T5):** I have not seen the movie yet, but my friend recommended it.

Table 4: Qualitative comparison of example translations from the G2T and T2G tasks. The fine-tuned model generates more fluent and grammatically complete sentences.

5 Discussion

Our experimental results provide compelling evidence that fine-tuning pre-trained language models (PLMs) is a superior strategy to training from scratch for bidirectional gloss translation. Across datasets, PLMs strongly outperform the baseline on G2T, and often on T2G as well—notably Llama 8B on ASLG-PC12—though some PLMs underperform the baseline on ASLG-PC12 T2G. This is demonstrated by concordant gains across BLEU, ROUGE-L, and METEOR metrics, alongside corresponding reductions in WER, as detailed in Tables 2 and 3.

5.1 The Decisive Advantage of Pre-trained Knowledge

A primary finding is the sheer magnitude of the improvement attributable to pre-training. On the challenging PHOENIX-14T dataset (G2T task), even the `T5-base` model achieves a BLEU-4 score of 22.73, a relative gain of roughly 74% over the 13.06 baseline. Larger or more sophisticated PLMs amplify this advantage, with `Llama 8B` reaching an impressive 29.92 BLEU-4.

This performance leap stems from the effective transfer of linguistic knowledge. As the qualitative examples in Table 4 illustrate, PLMs move beyond simple surface-level pattern matching. Compared to the telegraphic and grammatically incomplete outputs of the baseline, fine-tuned models successfully infer correct grammatical structure, inserting necessary function words, handling verb tenses, and producing far more natural and fluent sentences. This demonstrates that the models leverage their vast pre-trained knowledge of language, needing only to learn the mapping from glosses during fine-tuning.

5.2 The Bidirectional Bottleneck: Asymmetry in Translation

A critical insight from our bidirectional analysis is the significant asymmetry between the two translation directions. As shown in Figure 5 and in the detailed results in Table 2 and Table 3, we observe a stark performance asymmetry between the G2T and T2G directions across all models. Text-to-Gloss (T2G) translation is substantially more challenging than Gloss-to-Text (G2T). Across most models and datasets, we observe substantial BLEU reductions (often $\sim$30–60%) when reversing direction.

This difficulty arises because T2G requires the model to generate a syntactically rigid and often non-fluent sequence, which involves precise word deletion and reordering to match sign language grammar. This is an unnatural task for PLMs, whose pre-training objective is biased towards generating fluent, natural language. For instance, while `mBART` achieves a strong 25.58 BLEU-4 on the G2T task for PHOENIX-14T, its performance drops to just 12.10 for T2G. These findings confirm that in any practical bidirectional system, the T2G component is likely to be the primary performance bottleneck if not explicitly optimized with task-specific architectures or objectives. On T2G, a consistent counterexample appears: the from-scratch Transformer surpasses smaller pre-trained models on ASLG-PC12 dataset. This pattern supports a negative-transfer explanation, in which fluency-oriented pre-training conflicts with generating non-fluent, rule-like gloss sequences, while the baseline's neutral inductive bias learns the rigid mapping directly. Only very large models appear to mitigate this interference through additional capacity.

5.3 Architectural and Data Scale Considerations

Our results offer insights into the interplay between model architecture, pre-training objectives, and data conditions. Encoder-decoder models (T5, Flan-T5, mBART) prove highly competitive, especially on the smaller, real-world datasets like PHOENIX-14T and SIGNUM. Notably, the multilingual denoising pre-training of `mBART` appears to provide an advantageous inductive bias for the gloss-to-text mapping.

However, the decoder-only `Llama 8B` model excels where the data scale is largest, achieving the highest scores on the synthetic ASLG-PC12 dataset (83.10 BLEU-4 for G2T). This pattern suggests that while encoder-decoder architectures may be more data-efficient for learning the structured mapping from gloss to text, powerful decoder-only models can surpass them when sufficient parallel data is available to specialize to the task. Furthermore, the mixed results of instruction-tuning (Flan-T5 vs. T5) indicate that generic instruction-following priors do not always translate into downstream advantages for this highly structured translation task.

Finally, dataset characteristics clearly shape outcomes. The high scores on SIGNUM (up to 67.60 BLEU-4 G2T) highlight the effectiveness of PLMs on domain-specific data with controlled vocabularies. In contrast, PHOENIX-14T remains the most realistic and challenging benchmark, where our improvements represent substantial progress towards deployable, real-world systems.

5.4 Limitations and Future Work

Our evaluation relies primarily on automatic metrics and gloss-based representations, which do not capture non-manual markers and may not fully reflect end-user utility. Human evaluation, including DHH raters, should complement automatic metrics. From a modeling standpoint, `Llama 8B` raises compute and memory considerations; future work will investigate parameter-efficient tuning and knowledge distillation. Finally, closing the bidirectional gap likely requires objectives and architectures tailored for T2G (e.g., stronger constraints or structured decoding) and, longer term, integration with end-to-end video models that capture non-manual features.

6 Conclusion

We presented a comprehensive, controlled evaluation of pre-trained language models for bidirectional gloss translation across three distinct datasets. Our findings conclusively show that fine-tuning PLMs consistently and substantially outperforms training Transformers from scratch, with relative BLEU-4 gains on the G2T task ranging from roughly 74% (e.g., 13.06 $\rightarrow$ 22.73 on PHOENIX-14T with `T5-base`) to about 130% (13.06 $\rightarrow$ 29.92 with `Llama 8B`).

Our G2T results establish new state-of-the-art levels on all three benchmarks within our experimental setting, PHOENIX-14T (29.92 BLEU-4), SIGNUM (67.60), and ASLG-PC12 (83.10), demonstrating that transfer learning is a decisive enabler for this low-resource translation problem. The architectural analysis indicates that while encoder-decoder PLMs are highly competitive on smaller datasets, decoder-only LLMs can excel as data scale increases.

At the same time, our bidirectional study underscores a persistent asymmetry: Text-to-Gloss translation remains notably harder than Gloss-to-Text, with $\sim$30–60% BLEU reductions and elevated WER across datasets. Addressing this gap is a key avenue for future research, potentially requiring specialized objectives or constrained decoding.

Practically, these findings lower the barrier to building effective gloss translation systems. Strong models can be obtained via fine-tuning rather than costly training from scratch, making it feasible to extend SLT technology to additional sign languages and domains. We will release our code and fine-tuned checkpoints to support reproducibility and accelerate progress toward inclusive, deployable communication tools for the DHH community.

References

Abhimanyu Dubey Aaron Grattafiori. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa

Verhoef, and 1 others. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 16–31.

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)*, pages 7784–7793.

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, Minneapolis, Minnesota.

Pooya Fayyazsanavi, Antonios Anastasopoulos, and Jana Kosecka. 2024. Gloss2Text: Sign language gloss translation using LLMs and semantically aware label smoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16162–16171, Miami, Florida, USA. Association for Computational Linguistics.

Arshia Kermani, Veronica Perez-Rosas, and Vangelis Metsis. 2025. A systematic evaluation of llm strategies for mental health text analysis: Fine-tuning vs. prompt engineering vs. rag. *Preprint*, arXiv:2503.24307.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online.

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

Nazanin Mahjourian and Vinh Nguyen. 2025. Sanitizing manufacturing dataset labels using vision-language models. *arXiv preprint arXiv:2506.23465*.

Mohsen Mohammadagha, Israel Tshitenge, and Ifetilayo Adebambo. 2025. State-of-the-art machine learning techniques in sentiment analysis for social media: L'état de l'art des techniques d'apprentissage automatique en analyse de sentiment pour les médias sociaux.

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

Achraf Othman and Mohamed Jemni. 2012. English-ASL gloss parallel corpus 2012: ASLG-PC12. In *Proceedings of the LREC2012 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, pages 151–154, Istanbul, Turkey. European Language Resources Association (ELRA).

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

E.V. Pikoulis, A. Bifis, M. Trigka, C. Constantinopoulos, and D. Kosmopoulos. 2022. Context-aware automatic sign language video transcription in psychiatric interviews. *Sensors*, 22(7).

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Wendy Sandler and Diane Lillo-Martin. 2006. *Sign language and linguistic universals*. Cambridge University Press.

Mohammad Jalili Torkamani, Negin Mahmoudi, and Kiana Kiashemshaki. 2025. Llm-driven adaptive 6g-ready wireless body area networks: Survey and framework. *arXiv preprint arXiv:2508.08535*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Ulrich von Agris and Karl-Friedrich Kraiss. 2010. SIGNUM database: Video corpus for signer-independent continuous sign language recognition. In *Proceedings of the LREC2010 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 243–246, Valletta, Malta. European Language Resources Association (ELRA).

Kayo Yin and Jesse Read. 2020. Attention is all you sign: Sign language translation with transformers. In *Sign Language Recognition, Translation and Production (SLRTP) Workshop - Extended Abstracts*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas.