

## Supplemental Material

### Bootstrapping Sign Language Annotations with Sign Language Models

#### 1. LLM Prompts

In the paper we share results using three LLM prompts: (1) K-Shot English to ASL Gloss, (2) Fingerspelling: Error correction, and (3) Oracle back-translation: Manual Gloss Annotations to English. The full prompts are below.

##### **Prompt: Direct Translation (K-Shot English to Gloss)**

You are an expert American Sign Language (ASL) translator. Your task is to translate English sentences into ASL glosses using standard ASL linguistic notation. You may be asked to generate more than one candidate translation.

##### **OUTPUT FORMAT:**

- Provide ONLY the ASL gloss translation
- No explanations, notes, definitions, or commentary
- No punctuation marks
- If asked to translate multiple translations, each translation should be enumerated with the dictionary structure below.
- Use ALL CAPS for signs

##### **ASL GLOSS RULES:**

###### **1. NUMBERS:**

- Write numbers 1-9 as digits: 1, 2, 3, etc.
- Numbers 10+ use conceptual signing: "100" → "1 HUNDRED", "25" → "2 5" or "TWENTY-FIVE"
- Years: Use full digits "1998" or conceptual "NINETEEN NINETY-EIGHT"
- Ages: "AGE" + number
- Time: Use appropriate time markers (MORNING, AFTERNOON, etc.)

###### **2. GRAMMAR MARKERS:**

- Use hyphens for compound signs: SELF-CONTROL
- Fingerspelling: JOHN for names not having signs. Don't use "FS-" or have dashes between letters.
- Use classifiers when appropriate (e.g., CL:1(PPOINT) or CL:4(list))
- Prefer classifiers instead of index notation
- Use a special character instead of text (e.g., use "+" instead of "PLUS")

###### **3. ASL STRUCTURE:**

- Every sentence you generate should consider using a different grammatical structure.
- Sometimes follow ASL word order (typically Time-Subject-Object-Verb)

###### **4. COMMON CONVENTIONS:**

- Past tense: FINISH or time markers, not English -ed
- Questions: Use question markers and facial expressions

- Negation: Use NOT before the sign

EXAMPLE:

English: "I am happy"

Output:

“json

"1": "I AM HAPPY",

"2": "CL:1(point) AM HAPPY",

"3": "I VERY HAPPY",

"4": "ME AM HAPPY",

"5": "I SO HAPPY",

"6": "I REALLY HAPPY",

"7": "I AM GLAD",

"8": "I FEEL GOOD",

"9": "ME FEEL HAPPY",

"10": "I AM JOY"

Please generate {k} different gloss translations

NOW COMPLETE:

English: "{phrase}"

Output:

““

### **Prompt: Fingerspelling: Error correcting prompt**

Your task is to take a line of english text and fix typos. The text may contain a name, a physical address, a phone number, or a website address. First think about if it is a website, phone number, name, or home address. Then think about how it might be wrong and how to improve.

RULES:

- The input may be correct as-is or it may need to be fixed.
- Sometimes websites are only the terms in slashes like '/home/site/test/' or 'home-site-test'. Websites should have no spaces.
- Phone numbers generally have dashes between sets of digits.
- Do not remove full words or add new words.
- Do not abbreviate names. Do not add commas. Do not remove any numbers.
- Respond with a string response only with nothing else. Do not wrap with json. Do not add notes.

NOW COMPLETE:

English: "{phrase}"

Output:

### **Prompt: Oracle Experiment: Manual gloss annotations to English (back translation)**

You are an American Sign Language translator. Your task is to translate ASL glosses into complete English sentences. No notes, definitions, or anything other than glosses. Grammar should be in standard English word order. You may see other unicode symbols so translate those as you see fit. Words annotated with 'FS-' at the beginning are fingerspelled and should be replaced with the same word in English. Respond with a string response only with nothing else. Do not add notes.

NOW COMPLETE:  
 Glosses: "{phrase}"  
 Output:

## 2. Limitations of BLEU for Sign Language Translation

Traditional machine translation metrics like BLEU [1], which rely on n-gram overlap between predicted and reference translations, prove inadequate for evaluating sign language translation systems. BLEU’s focus on exact lexical matches fails to capture the semantic equivalence that is crucial in sign language translation, where multiple valid translations can express the same meaning using different vocabulary, word order, or levels of detail. This limitation is particularly pronounced in sign language translation due to the inherent differences between signed and spoken language modalities, including more flexible word order, frequent omission of function words, and the use of spatial and visual elements that may not have direct textual equivalents.

Our toy examples, shown in Table 1, demonstrate clear cases where BLEU scores are misleading or counterintuitive. All metrics range from 0 to 1 where higher is better. For instance, a translation that captures the core semantic meaning but uses synonyms or slightly different phrasing receives a low BLEU score despite being perfectly acceptable to human evaluators. Conversely, translations that maintain surface-level similarity through exact word matches but miss crucial semantic nuances can achieve artificially high BLEU scores while being rated poorly by human judges. These examples highlight BLEU’s inability to account for semantic equivalence, paraphrasing, and the contextual appropriateness that human evaluators naturally consider when assessing translation quality.

In contrast, learned metrics such as COMET [4] and BLEURT [5], which leverage pre-trained language models to capture semantic similarity rather than surface-level overlap, show significantly better alignment with human judgment in our evaluation. COMET’s ability to recognize semantic equivalence between different lexical realizations makes it particularly well-suited for sign language translation, where the goal is faithful meaning transfer rather than literal word-for-word correspondence. The superior correlation between COMET scores and human ratings across our test cases suggests that the field should move toward adopting these more sophisticated evaluation metrics.

ChrF [3] is also a popular machine translation metric and is a reasonable middle ground between BLEU and COMET. Unlike BLEU’s strict n-gram matching, ChrF operates at the character level and uses F-score rather than precision-based scoring, making it more robust to morphological variations and word order differences that are common in sign language translation. While ChrF still relies on surface-level string matching rather than semantic understanding, its character-based approach allows it to capture partial matches and morphological similarities that BLEU would miss entirely. In our evaluation, ChrF scores show moderate correlation with human judgment—better than BLEU but not as strong as COMET—reflecting its position as an improved lexical metric that nonetheless lacks the semantic awareness necessary for optimal sign language translation evaluation.

These findings have important implications for sign language translation research, as the choice of evaluation metric directly influences model development, comparison of different approaches, and assessment of progress in the field. We

Type	Phrase				
English Reference	I will go to the store tomorrow.	BLEU	ChrF	BLEURT	Comet <sub>22</sub>
ASL-like glosses	TOMORROW STORE TRAVEL I				
ASL-like back translation	Tomorrow I will travel to the store.	27.05	59.72	0.767	0.84
SEE-like glosses	I TRAVEL STORE TOMORROW				
SEE-like back translation	I will travel to the store tomorrow.	59.46	77.58	0.795	0.89
English Reference	The window was broken by the baseball.	BLEU	ChrF	BLEURT	Comet <sub>22</sub>
ASL-like glosses	BASEBALL HIT WINDOW, WINDOW BREAK				
ASL-like back translation	The baseball hit the window and the window broke.	6.27	54.30	0.681	0.79
SEE-like glosses	WINDOW BROKEN FROM BASEBALL				
SEE-like back translation	The window broke from the baseball	26.65	62.75	0.719	0.81

Table 1. Results across NLP metrics when back-translating an English reference with ASL-like glosses and Signed English (SEE)-like glosses. All back-translations accurately represent the English reference and thus are high quality, but scores for BLEU are very different, ChrF metrics are somewhat different, and both Comet and BLEURT are similar.

recommend that future work in sign language translation prioritize semantic-aware metrics like COMET over traditional n-gram based metrics, and encourage the development of specialized evaluation frameworks that account for the unique linguistic properties of sign languages and their translation into spoken language text.

### 3. Architecture Comparison: TCN vs. Conformer

We selected Temporal Convolutional Networks (TCNs) for their fixed receptive field, hypothesizing this would improve cross-dataset generalization compared to attention-based architectures. Post-submission experiments with a Conformer validate this choice. As shown in Table 2, while the Conformer achieves better in-domain results on FSBoard (which contains only fingerspelling), it overfits and performs significantly worse on out-of-domain data from ASL STEM Wiki and FLEURS-ASL, which contain a mixture of signing and fingerspelling. For a system designed to pseudo-annotate diverse real-world signing data, the TCN’s superior generalization is the preferred engineering choice.

Dataset	Conformer		TCN	
	All	Long	All	Long
FSBoard (CER ↓)	<b>4.9%</b>	–	7.3%	–
ASL STEM Wiki (AUC ↑)	0.70	0.79	<b>0.80</b>	<b>0.89</b>
FLEURS-ASL (AUC ↑)	0.69	0.73	<b>0.72</b>	<b>0.76</b>

Table 2. Fingerspelling results comparing Conformer and TCN architectures. FSBoard is in-domain (only fingerspelling); ASL STEM Wiki and FLEURS-ASL are out-of-domain (signing & fingerspelling). “Long” refers to words with  $\geq 3$  characters.

## 4. LLM Ablation Studies

### 4.1. Model Selection

We performed extensive experiments with different LLMs for gloss-to-English back-translation on ASL STEM Wiki. Table 3 shows BLEURT scores when translating from annotated ASL glosses to English. Larger models tend to perform better, but there is low variance between frontier models.

Model	BLEURT ↑
Gemma-3:12b	0.52
Gemma-3:27b	0.55
Claude Sonnet 4.0	0.59
Gemini-2.5-flash	0.60

Table 3. BLEURT scores for different LLMs on gloss-to-English back-translation using ASL STEM Wiki manual annotations.

### 4.2. K-Shot Prompting Ablation

There are many valid English-to-ASL translations, including some with very different grammar. We validated the K-shot approach by measuring ChrF when comparing manual annotations to the best-case LLM translation for  $K = 1$  to  $K = 10$ . As shown in Table 4, gloss-level ChrF improves from 47.7 ( $K = 1$ ) to 50.8 ( $K = 10$ ), with 67% of the improvement achieved by  $K = 4$ . This improvement is consistent for shorter and longer phrases.

$K$	Gloss-level ChrF ↑
1	47.7
4	49.8
10	50.8

Table 4. Effect of K-shot prompting on gloss-level ChrF, comparing manual annotations to the best-case LLM translation.

## 5. Summary of ASL STEM Wiki Interpretations

There were 16 unique signers across the videos annotated. The following summarizes the interpreters subjective assessments. The signers in ASL STEM Wiki can be generally separated into what appears to be native or native-like ASL users and second-language (L2) learners. Strong ASL users (Subjects 3, 4, 9, 11, 12, and 14) demonstrate proper ASL grammar structure, effective use of classifiers and depiction, clear fingerspelling, and appropriate use of non-manual markers. These signers also tend to avoid unnecessary fingerspelling when established signs exist and show fluent, natural signing patterns without hesitation. These signers, particularly Subjects 3 and 4, represent the gold standard for ASL output, with Subject 4 showing strategies typical of Certified Deaf Interpreters who aim for maximum comprehension across diverse audiences.

In contrast, several signers (Subjects 1, 6, 7, 10, 13, 15, and 16) exhibit what the Deaf community describes as a “hearing accent,” which are characteristics typical of second language (L2) ASL learners [2]. Such signers show varying degrees of English influence, from Subject 1’s heavy reliance on Signing Exact English (SEE) and fingerspelling, to Subject 10’s less precise fingerspelling style and sign initialization. These L2 signers typically follow English word order, have limited classifier use, and may hesitate or re-start mid-sentence while processing their interpretations. Some L2 signers also adopt textbook-style signing that lacks the natural flow of native users and employ unnecessary or less precise fingerspelling.

## References

- [1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. 3
- [2] Deborah Chen Pichler. Sources of handshake error in first-time signers of asl. *Deaf around the world: The impact of language*, pages 96–121, 2011. 5
- [3] Maja Popović. chrF: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395, 2015. 3
- [4] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, Nov. 2020. Association for Computational Linguistics. 3
- [5] Thibault Sellam, Dipanjan Das, and Ankur Parikh. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7881–7892, 2020. 3