Can Small Vision–Language Models Perform Sign Language Translation?

Anonymous ACL submission

Abstract

Vision-Language Models (VLMs) have shown strong generalization across multimodal tasks, but their capacity to handle *sign language translation (SLT)*—which requires fine-grained spatiotemporal reasoning and linguistic understanding—remains unclear. In this study, we evaluate whether *small-scale VLMs* ($\leq$3B parameters) can perform SLT effectively. We conduct supervised fine-tuning using multilingual sign language datasets—DGS, ASL, and ISL—adopting parameter-efficient LoRA tuning applied to the language decoder, while keeping the vision encoder frozen and allowing the connector to be trainable. To evaluate translation quality, we propose entity- and semantics-aware metrics tailored for SLT. We highlight the data imbalance issues present in the above widely used SLT datasets. Our analysis highlights the limitations in applying general-purpose VLMs to SLT, unlike their applicability in other tasks, and provides insights to inform future development of VLMs for SLP, which is essential for building inclusive AI applications.

1 Introduction

Sign language serves as a vital mode of communication for the deaf and hard-of-hearing (DHH) community, encompassing both manual components (e.g., hand gestures) and non-manual cues such as facial expressions, body posture, and eye gaze (Boyes Braem and Sutton-Spence, 2001). According to the World Health Organization[1], over 700 million people are expected to experience disabling hearing loss by 2050. To bridge the communication gap between DHH individuals and hearing populations, Sign Language Translation (SLT) broadly encompasses translation across modalities, including sign-to-text/speech, speech-to-sign, and sign-to-sign translation. In this work, we focus on the sign-to-text setting, where sign language videos are translated into written text. However, the diversity of sign languages, each with unique grammar and regional variations, alongside their rich spatio-temporal structure, poses substantial challenges and is, therefore, attracting increasing attention from the deep learning community. Indeed, researchers have strongly advocated the development of NLP tools for sign language understanding (Yin et al., 2021).

Unlike Sign Language Recognition (SLR), which predicts glosses or isolated labels, SLT requires models to capture the semantic, temporal, and syntactic intricacies of continuous signing. SLT methods typically fall into two categories: gloss-based and gloss-free. The former follows a two-step pipeline that first predicts glosses and then translates them into natural language (Camgoz et al., 2018, 2020; Zhou et al., 2021; Chen et al., 2022a; Zhou et al., 2022; Chen et al., 2022b), whereas the latter directly generates the translation bypassing the direct involvement of gloss annotation (Lin et al., 2023; Gong et al., 2024; Wong et al., 2024). Recent progress has been driven by transformer-based architectures (Vaswani et al., 2017) that take a visual sequence as input into a spatio-temporal encoder and generate a context-aware translation with a decoder which is essentially a large language model (LLM) (Gong et al., 2024; Jang et al., 2025; Wong et al., 2024).

Recent advancements in LLMs such as GPT-3 (Brown et al., 2020) and LLaMA (Touvron et al., 2023) demonstrated strong language understanding, reasoning, and multilingual translation by leveraging large-scale web corpora. Building on these capabilities, vision-language models (VLMs) like CLIP (Radford et al., 2021), BLIP (Li et al., 2022), LLaVA (Liu et al., 2023), and Qwen2.5-VL (Bai et al., 2025) aligned visual content with textual semantics and achieved state-of-the-art results across several multimodal tasks, demonstrating strong

[1] https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

temporal and visual reasoning – skills also essential for SLT. Although scaling model size has led to performance gains across many domains (Chowdhery et al., 2023; Alayrac et al., 2022; Liu et al., 2023), these large models often require considerable computational resources and are difficult to deploy in real-world or low-resource settings. This led the research interest in more efficient and scalable architectures.

Motivated by the advancement in both fields, we aim to investigate whether the multimodal or visual language models (Yin et al., 2024; Caffagni et al., 2024) can perform SLT effectively. In particular, we conduct supervised fine-tuning of off-the-shelf small-scale VLMs with a frozen visual model, trainable connector (modality aligner), and a LoRA tunable language model. Note that this fine-tuning methodology has shown a promising performance in several challenging tasks like radiology report generation (Chen et al., 2025; Kapadnis et al., 2024; He et al., 2025) and science question-answering (Kim et al., 2023). We also note that full fine-tuning of the VLM is computationally very expensive and infeasible with our infrastructure. Thus, our central research question is, *Can small-scale vision-language models, originally trained for generic video-language tasks, be effectively adapted for the specialized task of sign language translation?* This question is particularly challenging because SLT is significantly different from usual video understanding tasks like captioning and question-answering.

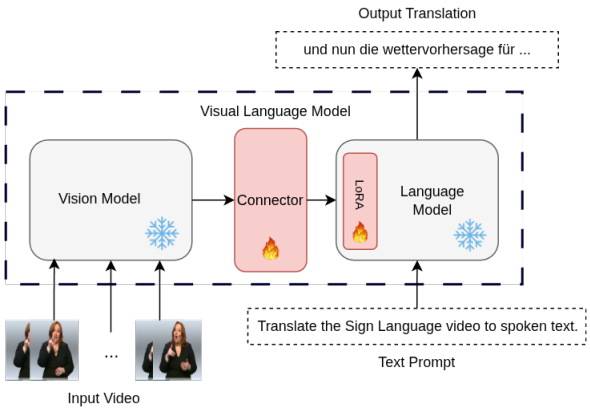

Figure 1: Overview of our vision-language model architecture for sign language translation.

In this work, we present the first systematic evaluation of general-purpose small VLMs for SLT, revealing that these models fall severely short of specialized SLT systems. Our key contributions are: (i) We conducted comprehensive experiments with four sign language datasets and three small VLMs across model sizes and fine-tuning types under resource-constrained settings where the vision encoder is frozen; (ii) Our experiments demonstrate consistent underperformance of VLMs in SLT; (iii) We uncover vocabulary biases in benchmark datasets that influence model outputs; and (iv) We introduce two targeted evaluation metrics, *translation soft-recall* and *psuedo-gloss entity-recall*, to better capture the semantic fidelity in SLT outputs. We will publicly release our code (upon acceptance of our manuscript).

2 Related Work

2.1 Sign Language Translation

Recent advances in SLT are driven by leveraging vision encoders and LLMs. Most of them target direct gloss-free translation by aligning visual features with frozen language models (Gong et al., 2024; Wong et al., 2024), supported by factorized adaptation (Chen et al., 2024), contextual reasoning (Jang et al., 2025), and video-language alignment pretraining (Jiao et al., 2025). Several studies also explored multilingual and continual learning settings (Yazdani et al., 2025; Zhang et al., 2024). Common datasets include PHOENIX-2014T (German) (Camgoz et al., 2018), How2Sign (American Sign Language) (Duarte et al., 2021), and OpenASL (ASL) (Shi et al., 2022).

2.2 Vision-Language Models

Following the success of LLMs like GPT-3 and LLaMA, researchers have attempted to incorporate multiple modalities beyond text into the *foundation models*. Multimodal large language models (MLLMs) and VLMs have emerged as their natural successor (Caffagni et al., 2024), and they support downstream tasks such as video captioning (Nadeem et al., 2024), video question answering (Zou et al., 2024; Xiao et al., 2024; Wang et al., 2025), and long video understanding (Ranasinghe et al., 2025), where cross-modal and temporal alignment is essential. These tasks can be achieved by zero or few-shot prompting. Fine-tuning, if necessary, is generally done on the LLM and the connector between the visual encoder and text decoder (Li et al., 2023; Yin et al., 2024). Recent research has also focused on developing smaller VLMs like SmolVLM2 (Marafioti et al., 2025), BLIP-2 (Li et al., 2023), TinyLLaVA (Zhou

et al., 2024), Qwen-2.5-VL (Bai et al., 2025), and InternVL3 (Zhu et al., 2025) which contain fewer than 3B parameters, which are attractive for resource-constrained and real-time deployments. However, it is unclear whether these models are suitable for SLT, and the papers that introduce these models do not mention if they have been trained or evaluated for sign language processing tasks despite their importance in creating inclusive AI systems. This paper aims to address that gap.

3 Method

3.1 Overview

Given a sign language video $V = [f_1, f_2, \cdots, f_T]$ with a sequence of T frames, the goal of SLT is to generate a spoken language sentence $Y = [w_1, w_2, \cdots, w_S]$, where $S \ll T$. Unlike recent works that design specialized transformer-based models to convert V to Y, we explore whether off-the-shelf VLMs, supporting video inputs, are suitable for this purpose since they are already pre-trained on large visual-text corpora. We choose small VLMs as they are more resource-friendly. Our model, shown in Figure 1, consists of a frozen vision encoder, a trainable connector which is a multi-layered perceptron, and a language model (LM). While the vision model extracts frame-level features from V, the connector projects them into the language embedding space, and the LM generates the final spoken language text. The connector is trained and the LM is fine-tuned.

3.2 VLM Selection.

We have selected three recent VLMs that strike a balance between efficiency and performance. The first is **SmolVLM2** (Marafioti et al., 2025), a 2.2B parameter model designed for edge deployment, known for its lightweight architecture and strong multimodal capabilities. SmolVLM2 is built on Idefics3 (Laurençon et al., 2024), comprising a 27-layer SigLIP-SO400M (Zhai et al., 2023) vision encoder and a 24-layer LLaMA 3.1 (Grattafiori et al., 2024) language model, connected via an MLP-based modality bridge. SmolVLM2 2.2B model achieves state-of-the-art performance on many image and video understanding tasks among open-weights models of similar scale, while being more resource-efficient (Marafioti et al., 2025). The second model we choose is **Qwen2.5-VL** (Bai et al., 2025), a 3B parameter architecture that integrates a 32-layer Vision Transformer (ViT) with efficient windowed attention and a 36-layer Qwen2.5 language model. Like SmolVLM2, it employs an MLP-based connector to merge visual and linguistic representations effectively. Finally, we include **InternVL3-2B**, an efficient open-source VLM trained with a native multimodal pre-training strategy that jointly learns visual and linguistic representations from mixed image–text, video–text, and text-only data. It pairs a 24-layer InternViT-300M-448px-V2.5 vision encoder with a 28-layer Qwen2.5-1.5B language model via a lightweight MLP connector. This unified training avoids post-hoc modality alignment and yields stronger temporal–semantic grounding, while Variable Visual Position Encoding (V2PE) (Ge et al., 2025) enables efficient handling of long multimodal and video contexts. Despite its compact scale, InternVL3-2B achieves competitive performance on video and long-context benchmarks, making it well suited for efficient video-based modeling.

3.3 Frame Sampling Strategy

Sign language videos vary widely in duration, with sequences ranging from a few frames to several hundred. However, the VLMs used in this study, namely, SmolVLM2, Qwen2.5-VL, and InternVL3 have a short context size (e.g., 8K tokens for SmolVLM2) which constrains the number of video frames to, say, T_{max}, depending on the size of the input image patches and the model's internal architecture (e.g., 64 for SmolVLM2 with default settings). The model features are shown in Table 1.

To ensure that each video is processed efficiently within this limit, we judiciously select a frame count $T_{\max}$ based on the model's capacity and the video length distribution. In particular, given a video with T total frames, where $T > T_{\max}$, we divide the sequence into $T_{\max}$ equal-length intervals and select one representative frame per interval. The representative frame f_i in each interval is selected deterministically using Eq (1):

$$f_i = \mathbf{frames}\left[\left\lfloor \frac{(i+1)\cdot T}{T_{\max}} \right\rfloor - 1\right], \qquad (1)$$

where $i = (0, 1, \ldots, T_{\max} - 1)$ and $\mathbf{frames}[\cdot]$ denotes frames whose indices appear in $[\cdot]$. This ensures a uniform spread of selected frames across the entire video duration. If $T \leq T_{\max}$, we retain all frames without modification.

Note that we do not use random skipping as it may drop critical frames, disrupting semantic con-

tinuity. The final output is a list of $T_{\max}$ or fewer frames, sampled in a way that balances global coverage and local coherence. Our design choice ensures fair and consistent evaluation across variable-length sign language videos while adhering to the strict frame limits imposed by our chosen VLMs.

3.4 Finetuning VLMs

The language models were tuned using the parameter-efficient Low-Rank Adaptation (LoRA) method (Hu et al., 2022), which allowed us to update only a small subset of parameters while maintaining the general linguistic capabilities of the models. We observed consistent performance improvements with the use of LoRA over full fine-tuning (see Table 9), particularly in resource-constrained settings. Note that the encoder was kept frozen while the connector was trained along with the LM.

Model	No. of Parameters	Vision Model (No. of layers)	Language Model (No. of layers)	Input Context Length
SmolVLM2	2.2B	SigLIP (27)	LLaMA-3.1 (24)	8K
Qwen2.5-VL	3B	ViT (32)	Qwen2.5LM (36)	32K
InternVL3	2B	InternViT-300M (24)	Qwen2.5-1.5B (28)	32K

Table 1: Overview of Selected VLM Architectures.

4 Experiments

4.1 Datasets

We evaluated our approach on four publicly available sign language translation datasets spanning German, American, and Indian sign languages. While some of these datasets included intermediate gloss annotations (e.g., PHOENIX-2014T), our training does not rely on gloss supervision. Instead, we focus on learning a direct mapping from sign language videos to natural language text.

German Sign Language (DGS): The RWTH-PHOENIX-Weather 2014T dataset (Camgoz et al., 2018) serves as a standard benchmark for DGS translation. It consists of 7,096 training, 519 validation, and 642 test samples aligned with German sentences. With a vocabulary of 2,887 words, it captures weather forecast scenarios performed by professional interpreters on television.

American Sign Language (ASL): OpenASL (Shi et al., 2022) is a large-scale ASL dataset collected from online videos across a wide range of topics. It includes 98,417 video-sentence pairs, featuring over 200 signers, of which 966 samples are reserved for validation and 975 for testing. We also use the How2Sign (Duarte et al., 2021) dataset, which provides around 120 hours of ASL content derived from instructional videos, with 31,128 training, 1,741 validation, and 2,322 test samples. It includes multi-view videos and additional modalities such as speech, keypoints, and depth maps.

Indian Sign Language (ISL): iSign (Joshi et al., 2024) is a recently introduced large-scale dataset for Indian Sign Language, offering over 127K sentence-aligned signing videos from diverse, real-world contexts. Due to its scale, we carefully select a subset of 50K examples: 43K for training and 3.5K each for validation and test, preserving linguistic diversity while ensuring that the sample size is not too large to train our models. More details of the selection process are provided in Appendix A.

Dataset	Total videos	Avg. frame count	Frame count > 300
PHOENIX-2014T	7,096	116.59	23
How2Sign	31,047	162.76	3,772
OpenASL	97,233	205.25	24,752
iSign	127,237	209.08	23,767

Table 2: Video-level statistics across different datasets used in our experiments. We report the total number of videos, the average number of frames per video, and the number of videos with more than 300 frames.

4.2 Evaluation Metrics

Standard Metrics. To evaluate the translation quality of sign language outputs, we have used standard metrics widely adopted in the machine translation literature: BLEU (Papineni et al., 2002) and ROUGE-L (Lin and Och, 2004). BLEU captures n-gram precision by comparing predicted translations to ground-truth references, and we reported scores from BLEU-1 through BLEU-4 using the SacreBLEU[2] implementation. ROUGE-L measures the longest common subsequence between the prediction and reference; it captures how well the predicted sentence preserves the sequence of the target, without requiring a predefined n-gram length. We additionally report BLEURT (Sellam et al., 2020), a learned evaluation metric based on BERT that correlates more strongly with human judgment by modeling semantic adequacy and fluency, even in low-resource and out-of-distribution settings..

[2] https://github.com/mjpost/sacrebleu

Semantic Fidelity in SLT. Given the generative properties of LMs, it is possible that the LM's output does not match the ground-truth translation but is still semantically correct. To capture this aspect, we introduce two metrics: *Translation Soft Recall* (TSR) and *Pseudo-Gloss Entity Recall* (PGER), which are adaptations of Heading Soft Recall (Fränti and Mariescu-Istodor, 2023) and Heading Entity Recall (Shao et al., 2024), respectively. They are defined in Eqs. (2), (3), (4). The former computes semantic similarity between predicted and reference sentences using cosine similarity of embeddings from the multilingual Sentence Transformer model `paraphrase-multilingual-MiniLM-L12-v2`[3]. This captures meaning beyond exact wording and can better handle paraphrase variations. The PGER metric is motivated by the observation that signers often emphasize key content words such as nouns, verbs, and adjectives. We extract these entities using FLAIR (Akbik et al., 2019) part-of-speech tagging and compute their overlap between prediction and reference, treating them as proxies for underlying glosses.

$$\mathcal{C}(Z) = \sum_{i=1}^{|Z|} \frac{1}{\sum_{j=1}^{|Z|} \mathrm{Sim}(Z_i, Z_j)} \tag{2}$$

$$\mathrm{TSR}(Y, \hat{Y}) = \frac{\mathcal{C}(Y) + \mathcal{C}(\hat{Y}) - \mathcal{C}(Y \cup \hat{Y})}{\mathcal{C}(\hat{Y})} \tag{3}$$

$$\mathrm{PGER}(Y, \hat{Y}) = \frac{|\,\mathrm{Ent}(Y) \cap \mathrm{Ent}(\hat{Y})|}{|\,\mathrm{Ent}(Y)|} \tag{4}$$

Here, Y and $\hat{Y}$ represent the ground truth and predicted translations, respectively. $Ent(\cdot)$ denotes the set of entities extracted from a translation. $\mathcal{C}(Z)$ refers to the soft cardinality of the set Z (Fränti and Mariescu-Istodor, 2023). The function $Sim(\cdot, \cdot)$ computes the cosine similarity between the embeddings of word Z_i and Z_j within the same translation. These metrics quantify how well a model preserves meaning while using words different from those in the golden output. To our knowledge, this is the first application of such entity- and semantics-aware metrics in SLT, providing deeper insights into translation faithfulness.

4.3 Implementation Details

Our implementation was done on the PyTorch framework, and experiments were conducted using a single NVIDIA A100 GPU. A more detailed system specification is provided in Appendix G.

Network Details. We fine-tuned three open-source VLMS from Hugging Face repository: **SmolVLM2-2.2B-Instruct**[4], **Qwen2.5-VL-3B-Instruct**[5], and **InternVL3-2B**[6]. Both models were trained for **10 epochs**, using sampled video inputs of length $T_{\max} = 64$ frames. Frame statistics for all datasets are detailed in Table 2. We froze the vision encoder in both cases and kept a lightweight connector module fully trainable. We employed LoRA-based parameter-efficient fine-tuning on the attention and MLP projection layers of all language models. Using default configurations, **SmolVLM2** and **Qwen2.5-VL-3B** were fine-tuned with a LoRA rank $r = 64$ and scaling factor $\alpha = 64$, with dropout rates of 0.1 and 0.05, respectively, while **InternVL3-2B** used a default rank of $r = 32$, $\alpha = 64$, and no dropout.

Training and Inference Details. We trained all models using the Hugging Face `Trainer`, with FlashAttention-2 enabled and computations in `bfloat16`. Optimization was performed using a standard Transformer training setup with AdamW, a per-device batch size of 1, and gradient accumulation over 4 steps. **SmolVLM2** was trained with a fixed learning rate of 1×10^{-4}, while **Qwen2.5-VL-3B** and **InternVL3-2B** used a cosine learning-rate schedule with peak learning rates of 1×10^{-5} and 2×10^{-5}, respectively. During inference, we used greedy decoding with a maximum generation length of 128 tokens for all models. The prompt used for fine-tuning and inference is presented in Appendix E.

4.4 Results

Table 3 reports the evaluation results using BLEU (1–4) and ROUGE-L scores on the validation (Dev) and test sets across four sign language datasets for the three evaluated models. Overall, InternVL3 and SmolVLM2 substantially outperform Qwen2.5-VL across all datasets. On PHOENIX-2014T, InternVL3 achieves the highest BLEU-4 score on the Dev set (8.51), while SmolVLM2 attains the best Test BLEU-4 score (8.44), significantly exceeding Qwen2.5-VL (2.78). Similar trends are observed on How2Sign, OpenASL, and iSign, where InternVL3 consistently produces the strongest BLEU

[3] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

[4] https://huggingface.co/HuggingFaceTB/SmolVLM2-2.2B-Instruct

[5] https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct

[6] https://huggingface.co/OpenGVLab/InternVL3-2B

Models	DEV					Test				
	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE-L	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE-L
RWTH-PHOENIX-Weather 2014T (PHOENIX-2014T) (German Sign Language)										
SmolVLM2-2.2B	25.53	15.51	10.11	7.80	25.91	24.15	14.89	**10.16**	**8.44**	24.51
Qwen2.5-VL-3B	7.62	3.25	2.98	2.56	9.98	7.87	3.43	3.09	2.78	9.79
InternVL3-2B	**27.32**	**16.78**	**10.96**	**8.51**	**27.01**	**25.97**	**15.65**	9.93	8.00	**26.49**
How2Sign (American Sign Language)										
SmolVLM2-2.2B	6.83	1.46	0.23	0.04	8.38	6.84	1.56	0.28	0.06	8.38
Qwen2.5-VL-3B	2.45	0.74	0.21	0.01	4.83	2.80	0.75	0.18	0.02	5.05
InternVL3-2B	**8.66**	**1.91**	**0.45**	**0.15**	**9.08**	**8.15**	**1.77**	**0.42**	**0.11**	**8.55**
OpenASL (American Sign Language)										
SmolVLM2-2.2B	7.64	1.91	1.05	0.79	8.47	7.06	1.20	0.60	0.41	7.78
Qwen2.5-VL-3B	3.24	0.30	0.11	0.01	4.31	3.52	0.29	0.08	0.07	4.46
InternVL3-2B	**13.47**	**6.95**	**3.93**	**2.70**	**14.50**	**13.67**	**7.12**	**4.27**	**2.96**	**14.66**
iSign (Indian Sign Language)										
SmolVLM2-2.2B	7.47	1.27	0.60	0.39	8.09	7.59	1.37	0.65	0.44	8.21
Qwen2.5-VL-3B	2.48	0.27	0.15	0.13	3.85	2.51	0.33	0.22	0.18	3.87
InternVL3-2B	**8.26**	**1.60**	**0.78**	**0.50**	**9.01**	**8.56**	**1.86**	**0.97**	**0.68**	**9.20**

Table 3: Sign language translation results across four datasets in German (RWTH-PHOENIX-Weather 2014T), American (How2Sign, OpenASL), and Indian (iSign) sign languages. We report BLEU (1–4) and ROUGE-L scores for both validation (Dev) and test sets using SmolVLM2, Qwen2.5-VL, and InternVL3.

and ROUGE-L scores, followed by SmolVLM2.

Model	PHOENIX-2014T	How2Sign	OpenASL	iSign
SmolVLM2	0.36	0.2831	0.2336	0.2685
Qwen2.5-VL	0.1507	0.2476	0.1909	0.2144
InternVL3-2B	**0.3981**	**0.2905**	**0.3145**	**0.2965**

Table 4: BLEURT Scores on the Test Sets of the Evaluated Datasets

Table 4 further reports BLEURT scores on the test sets, reflecting semantic similarity between predictions and references. InternVL3 achieves the highest BLEURT scores across all datasets, with SmolVLM2 ranking second and Qwen2.5-VL consistently trailing, confirming the overall trends observed in BLEU and ROUGE-L evaluations.

To assess translation quality at a deeper semantic level, we employ Pseudo-Gloss Entity Recall (PGER) and Translation Soft Recall (TSR). As shown in Tables 5 and 6, InternVL3 consistently achieves the highest scores across all datasets on both metrics, indicating stronger coverage of key content words and improved semantic recall. SmolVLM2 follows closely behind, while Qwen2.5-VL yields substantially lower scores in most cases.

Overall, both InternVL3 and SmolVLM2 outperform Qwen2.5-VL across datasets on lexical- and semantic-level evaluations. However, despite these relative gains, the BLEU and ROUGE-L scores on How2Sign, OpenASL, and iSign remain very low, suggesting that the generated translations for ASL and ISL datasets are still of limited quality. Consequently, we focus our subsequent analysis primarily on PHOENIX-2014T, where translation performance is comparatively more reliable.

Model	PHOENIX -2014T	How2Sign	OpenASL	iSign
SmolVLM2	0.1372	0.0678	0.0318	0.0190
Qwen2.5-VL	0.0221	0.0428	0.0128	0.0143
InternVL3	**0.1545**	**0.0547**	**0.1044**	**0.0226**

Table 5: Pseudo-Gloss Entity Recall (PGER) based on key content words (noun, verb, adjective) on test sets. Higher is better.

4.4.1 Qualitative Analysis of Translation Results

We examine a few cases from PHOENIX-2014T test set to understand the types of errors committed by SmolVLM2. The cases displayed in Table 7 reveals two recurring types of translation errors. First, the model frequently failed at entity ground-

Model	PHOENIX -2014T	How2Sign	OpenASL	iSign
SmolVLM2	0.7422	0.3102	0.2538	0.2611
Qwen2.5-VL	0.4857	0.3091	0.179	0.1195
InternVL3	**0.7576**	**0.3425**	**0.3753**	**0.2721**

Table 6: Translation Soft Recall (TSR) scores on test sets across datasets. Higher is better.

ing, for example, misidentifying specific dates or days (e.g., predicting *May 30th* instead of the correct *October 15th*). Second, it often hallucinated semantically plausible but incorrect content—such as adding weather details like "rain in the north," that were absent in the reference. These patterns indicate that translation of the SmolVLM2 frequently lacked precise alignment between the visual input and the target linguistic output. Additional results on various datasets are provided in Appendix B while a more detailed breakdown focused on noun-phrase-level performance and qualitative examples appear in Appendix C.

Ground Truth:	und nun die wettervorhersage für morgen freitag den fünfzehnten oktober
English Translation	And now the weather forecast for tomorrow, Friday, October 15th.
SmolVLM2:	und nun die wettervorhersage für morgen freitag den dreißigsten mai
English Translation	And now the weather forecast for tomorrow, Friday, May 30th.
Qwen2.5-VL:	und nun die neue richtung
English Translation	And now the new direction.
Ground Truth:	und nun die wettervorhersage für morgen donnerstag den fünfzehnten oktober
English Translation	And now the weather forecast for tomorrow, Thursday, October 15th.
SmolVLM2:	und nun die wettervorhersage für morgen montag den achtundzwanzigsten oktober
English Translation	And now the weather forecast for tomorrow, Monday, October 28th.
Qwen2.5-VL:	und nun die wettervorhersage für morgen montag den neunzehnten oktober
English Translation	And now the weather forecast for tomorrow, Monday, October 19th.
Ground Truth:	am freitag ist es in der nähe der küsten und ganz im süden teilweise stark bewölkt
English Translation	On Friday, it will be partly cloudy near the coast and in the far south.
SmolVLM2:	am freitag ist es im süden und osten weiter freundlich im norden regnet es gebietsweise
English Translation	On Friday, it will remain pleasant in the south and east, with rain in some areas in the north.
Qwen2.5-VL:	und jetzt wünsche ich ihnen noch einen schönen abend
English Translation	and now I wish you a nice evening.

Table 7: Translation examples from the PHOENIX-2014T test set. Errors are shown in red and corresponding ground truth in blue.

4.4.2 Vocabulary Bias and Overfitting Tendencies

We have observed that some words are very frequently present in the generated translation, and they often appear in place of the right word in the output. So we ask: Does the training set have a specific bias towards some set of words? To investigate it, we select the PHOENIX-2014T dataset and pick the top ten most frequent words and noun phrases in the predicted translations from the SmolVLM2 model. While we notice several high-frequency tokens (e.g., `und`, `es`, `im`) appeared consistently across all sets, words such as `süden` and `norden` were notably overrepresented in predictions despite being much less frequent in the test set, and they were also high-frequency words in the training set. This reflects a tendency of the model to overgenerate words that were more prominent in the training distribution. Similar trends were also observed in the noun phrase results. A more detailed comparison of vocabulary distributions across the predicted, test, and training sets is provided in Appendix D.

These tendencies of biasness motivated us to carry out a vocabulary analysis across the training set of all four datasets and the frequency of each unique word; details of the analysis are presented in Table 2. All datasets demonstrated a highly skewed distribution, with the majority of tokens occurring fewer than ten times. In particular, over 70% of the vocabulary in OpenASL and iSign fell into this low-frequency range. This long-tailed nature of the data likely contributed to the model's generation of high-frequency terms, impacting its ability to produce rare or diverse words during inference.

5 How do these results compare to SOTA?

Our findings show that small-scale VLMs, even with their LMs fine-tuned on sign language data, fail to achieve competitive performance on standard sign language translation benchmarks. The SOTA methods show a BLEU-4 score of 24.32 in the test set of the PHOENIX-2014T dataset, and 10.11 BLEU-4 on the test set of the How2Sign dataset (Hwang et al., 2025). While OpenASL is less explored due to its scale, the BLEU-4 score on its test set is 7.06 (Lin et al., 2023). iSign is a more recent dataset that has a SOTA BLEU-4 of 1.47 (Joshi et al., 2024). These numbers are *substantially higher* than the scores we obtained. We observe that the SOTA models employ very specialized architectures, though mostly based on transformers, that usually rely on training the vision encoder using pseudo-gloss (Lin et al., 2023; Gong et al., 2024) or visual-language pretraining to adapt the visual encoder to the complexities of sign language (Zhou et al., 2023) or designing modules to precisely capture spatial configurations and motion dynamics present in sign language (Hwang et al.,

2025). Clearly, general-purpose vision-language foundation models lack the inductive biases and task-specific training necessary to capture the spatiotemporal complexities of sign language, and this cannot be addressed simply by training the LM. This is likely because the models are exposed to little or no sign language data during pretraining, and is also not instruction-tuned on sign language processing tasks.

Dataset	No. of Tokens	No. of Unique Token	Token Freq.<10	Token Freq >100
PHOENIX-2014T	99,081	2,887	72.1 %	6.3 %
How2Sign	41,421	3,577	86.8 %	1.8 %
OpenASL	13,71,021	29,197	73.1 %	5.3 %
iSign	13,20,855	33,498	75.5 %	4.5 %

Table 8: Vocabulary statistics of different sign language datasets, including total token counts, number of unique tokens, and the distribution of rare (frequency < 10) and frequent (frequency > 100) tokens.

6 Ablation Study

Model Parameter Selection and Fine-tuning Strategy: We investigated three scales of the SmolVLM2 model (Marafioti et al., 2025)—256M, 500M, and 2.2B parameters—across various fine-tuning strategies. Full fine-tuning of the 256M model yielded poor performance and was therefore not pursued further. Due to resource constraints, full fine-tuning of the 500M model was impractical, so we explored partial strategies: i) tuning only the vision or language models, ii) sequentially tuning both, with the connector being trained with the language model (LM) and iii) applying LoRA to only LM while keeping the vision encoder frozen and fully training the connector.

The nature of the experimental results in the third category of experiments of 500M, we also scaled up to a 2.2B model with a similar strategy. This setup yielded the best overall results (Table 9), demonstrating that targeted fine-tuning of larger models led to significant performance gains without incurring the cost of full training.

Effect of Frame Length on Translation Quality: To maintain compatibility with SmolVLM2's architectural constraints—specifically, a maximum of 64 frames at 384p resolution[7]—we standardized all experiments to 64-frame video inputs. To assess whether longer sequences improved translation quality, we fine-tuned Qwen2.5-VL (Bai et al., 2025) with a context length of 256 frames on the PHOENIX-2014T dataset.

Results in Table 10 show that increasing the frame count from 64 to 256 consistently reduced performance, suggesting that the model is not capable of accurately processing longer videos. Based on these findings, we adopted 64-frame inputs for all models to balance translation quality with computational efficiency. Additional studies on InternVL3 are presented in Appendix F.

Model Size	Fine-tuning Strategy	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE-L
256M	Full	11.2801	3.1971	1.3475	0.9809	11.5488
500M	LM only	21.1849	11.3004	7.0104	5.7513	20.8966
500M	Vision model only	20.0229	9.8091	6.1418	4.9888	20.6542
500M	Vision then LM	22.3609	12.0529	7.8071	6.0799	21.9553
500M	LoRA for LM	**23.1022**	12.7053	7.8365	6.1264	**23.4076**
2.2 B	LoRA for LM	22.2051	**12.7622**	**8.2799**	**6.5134**	22.5206

Table 9: Ablation study comparing different SmolVLM2 model sizes and fine-tuning strategies on the PHOENIX-2014T test set (after 5 epochs).

Max Frames	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE-L
64	**7.8711**	**3.4272**	**3.0888**	**2.784**	**9.7947**
256	6.4473	2.2679	2.0075	1.7449	7.4508

Table 10: Impact of Input Frame Length on Translation Performance for Qwen2.5-VL Fine-Tuned on PHOENIX-2014T

7 Conclusion

In this work, we analyse the relevance of off-the-shelf VLMs on SLT task. Our findings show that the direct integration of VLM with fine-tuning only the language model (and not the vision encoder), which shows promising results in other domains, fails to produce satisfactory performance in the SLT task. This highlights a significant limitation of foundation models when it comes to the development of inclusive AI tools and frameworks. In particular, it is necessary to incorporate sign language understanding objectives in the pretraining and instruction-tuning phases of these models so that they can be used with zero or few-shot prompting for sign language translation and other related tasks for the more common sign languages, and can be easily adapted to various other sign languages with limited fine-tuning.

[7] https://huggingface.co/HuggingFaceTB/SmolVLM2-2.2B-Instruct/blob/main/preprocessor_config.json

Acknowledgments

We have used the GenAI tool ChatGPT for minor English editing at some places in the manuscript. However, no technical content in the manuscript was generated using ChatGPT or any other AI tool.

Limitations

Our architectural choices, including the video context length, were constrained by available computational resources and model input limitations. To reduce training cost, we froze the vision encoder during fine-tuning. Although we analyze the severe data imbalance present in commonly used sign language translation datasets, addressing this issue is beyond the scope of the current work. Our experiments are further limited to three small vision–language models and four SLT datasets, which may restrict the generality of our findings. Finally, we do not conduct a comprehensive societal or user-centered evaluation. While existing benchmarks enable fair comparisons, they are not fully representative of the diversity of sign language users. Moreover, real-world deployment would require careful bias analysis and community-informed evaluation.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

P Boyes Braem and RL Sutton-Spence. 2001. *The Hands Are The Head of The Mouth. The Mouth as Articulator in Sign Languages*. Hamburg: Signum Press.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The revolution of multimodal large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13590–13618.

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)*.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. 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 (CVPR)*.

Xupeng Chen, Zhixin Lai, Kangrui Ruan, Shichu Chen, Jiaxiang Liu, and Zuozhu Liu. 2025. R-LLAVA: IMPROVING MED-VQA UNDERSTANDING THROUGH VISUAL REGION OF INTEREST. In *ICLR 2025 Workshop on Human-AI Coevolution*.

Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022a. A Simple Multi-Modality Transfer Learning Baseline for Sign Language Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5120–5130.

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie LIU, and Brian Mak. 2022b. Two-Stream Network for Sign Language Recognition and Translation. In *Advances in Neural Information Processing Systems*, volume 35, pages 17043–17056. Curran Associates, Inc.

Zhigang Chen, Benjia Zhou, Jun Li, Jun Wan, Zhen Lei, Ning Jiang, Quan Lu, and Guoqing Zhao. 2024. Factorized Learning Assisted with Large Language Model for Gloss-free Sign Language Translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7071–7081. ELRA and ICCL.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, and 48 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

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)*.

Pasi Fränti and Radu Mariescu-Istodor. 2023. Soft precision and recall. *Pattern Recognition Letters*, 167:115–121.

Junqi Ge, Ziyi Chen, Jintao Lin, Jinguo Zhu, Xihui Liu, Jifeng Dai, and Xizhou Zhu. 2025. V2PE: Improving Multimodal Long-Context Capability of Vision-Language Models with Variable Visual Position Encoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21070–21084.

Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. 2024. LLMs are Good Sign Language Translators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18362–18372.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jinlong He, Pengfei Li, Gang Liu, Genrong He, Zhaolin Chen, and Shenjun Zhong. 2025. PeFoMed: Parameter Efficient Fine-tuning of Multimodal Large Language Models for Medical Imaging. *Preprint*, arXiv:2401.02797.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Eui Jun Hwang, Sukmin Cho, Junmyeong Lee, and Jong C. Park. 2025. An Efficient Gloss-Free Sign Language Translation Using Spatial Configurations and Motion Dynamics with LLMs. 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 3901–3920. 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. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 8742–8752.

Peiqi Jiao, Yuecong Min, and Xilin Chen. 2025. Visual Alignment Pre-training for Sign Language Translation. In *Computer Vision – ECCV 2024*, pages 349–367. Springer Nature Switzerland.

Abhinav Joshi, Romit Mohanty, Mounika Kanakanti, Andesha Mangla, Sudeep Choudhary, Monali Barbate, and Ashutosh Modi. 2024. iSign: A Benchmark for Indian Sign Language Processing. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10827–10844. Association for Computational Linguistics.

Manav Kapadnis, Sohan Patnaik, Abhilash Nandy, Sourjyadip Ray, Pawan Goyal, and Debdoot Sheet. 2024. SERPENT-VLM : Self-Refining Radiology Report Generation Using Vision Language Models. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*.

Sungkyung Kim, Adam Lee, Junyoung Park, Sounho Chung, Jusang Oh, and Jay-Yoon Lee. 2023. Parameter-efficient fine-tuning of instructblip for visual reasoning tasks. In *Efficient Natural Language and Speech Processing Workshop at NeurIPS*, volume 2023.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. Building and better understanding vision-language models: insights and future directions. *Preprint*, arXiv:2408.12637.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612.

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. 2023. Gloss-Free End-to-End Sign Language Translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12904–12916. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, and 1 others. 2025. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*.

Asmar Nadeem, Faegheh Sardari, Robert Dawes, Syed Sameed Husain, Adrian Hilton, and Armin Mustafa. 2024. Narrativebridge: Enhancing video captioning with causal-temporal narrative. *arXiv preprint arXiv:2406.06499*.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Kanchana Ranasinghe, Xiang Li, Kumara Kahatapitiya, and Michael S Ryoo. 2025. Understanding long videos with multimodal language models. In *The Thirteenth International Conference on Learning Representations*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892. Association for Computational Linguistics.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024. Assisting in writing Wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6252–6278.

Bowen Shi, Diane Brentari, Gregory Shakhnarovich, and Karen Livescu. 2022. Open-Domain Sign Language Translation Learned from Online Video. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6365–6379, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

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.

Xijun Wang, Junbang Liang, Chun-Kai Wang, Kenan Deng, Yu Lou, Ming C. Lin, and Shan Yang. 2025. ViLA: Efficient Video-Language Alignment for Video Question Answering. In *Computer Vision – ECCV 2024*, pages 186–204, Cham. Springer Nature Switzerland.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2024. Sign2GPT: Leveraging Large Language Models for Gloss-Free Sign Language Translation. In *The Twelfth International Conference on Learning Representations*.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. 2024. Can I Trust Your Answer? Visually Grounded Video Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13204–13214.

Shakib Yazdani, Josef Van Genabith, and Cristina España-Bonet. 2025. Continual Learning in Multilingual Sign Language Translation. 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 10923–10938. Association for Computational Linguistics.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including Signed Languages in Natural Language Processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986.

Biao Zhang, Garrett Tanzer, and Orhan Firat. 2024. Scaling Sign Language Translation. In *Advances in Neural Information Processing Systems*, volume 37, pages 114018–114047. Curran Associates, Inc.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. 2024. TinyLLaVA: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*.

Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. 2023. Gloss-free sign language translation: Improving from visual-language pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20871–20881.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving Sign Language Translation With Monolingual Data by Sign Back-Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1325.

Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. 2022. Spatial-Temporal Multi-Cue Network for Sign Language Recognition and Translation. *IEEE Transactions on Multimedia*, 24:768–779.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

Bo Zou, Chao Yang, Yu Qiao, Chengbin Quan, and Youjian Zhao. 2024. Language-aware Visual Semantic Distillation for Video Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27113–27123.

Appendix

This appendix provides additional details and analyses that support our main findings. It includes insights into dataset preparation, translation behavior, grounding errors, vocabulary analysis, and resource usage.

Appendix Table of Contents

Figure 2: Frames sampled from a video with id *aTB_lu2Im8Y–129* showing no human presence from iSign dataset.

A Dataset Selection and Splitting Strategy for iSign

The iSign dataset contains noise at various places. For example, some videos do not have any human presence as illustrated in Figure 2, while some have no valid translation texts. To improve the dataset quality, we first applied two basic cleanup steps: removing samples without valid textual translations and filtering out videos with no visually detectable human presence using pose estimation.

To efficiently fine-tune a small VLM, we reduced our dataset from 126,000 to 50,000 video-text pairs. Rather than applying random sampling, we followed a vocabulary-aware strategy as follows. A vocabulary frequency analysis revealed that over 75% of the words in the dataset occurred fewer than 10 times. These *rare words* are essential for balancing the overall vocabulary spectrum, yet they also pose a challenge for model generalization. To strike a balance, we first selected up to 30,000 samples that contained at least one rare word. We then sampled an additional 20,000 examples from the rest of the dataset, which ensures both rare and frequent expressions are present in the data subset we used.

This method ensures that the model is trained on a wide vocabulary, especially low-frequency tokens, while the validation and test sets remain challenging but not overly skewed toward rare or unseen words. The result is a dataset that better supports robust and generalizable model training.

B Qualitative Analysis of Translation Results.

As shown in Tables 11 and 12, we provide a few representative translation outputs from the OpenASL and iSign datasets. While the overall translation quality across both models is poor, we include a mix of relatively better and worse examples to highlight common trends. In many cases, outputs fail to preserve critical entities or semantic meaning, underscoring the challenges of sign language translation in current settings. Although

SmolVLM2 tends to generate slightly better translation in some cases, the overall performance of both models shows poor quality translation.

Ground Truth:	I am Ed Bosson, and I am 67 years old and now retired.
SmolVLM2:	Hello, my name is E. Simmons and I am a retired school administrator.
Qwen2.5-VL:	I'm Ez Sussy.
Ground Truth:	We are also supporting efforts of DeafBlind people to develop hospital access because their needs are different from other deaf and hard of hearing people.
SmolVLM2:	The IDEA has been in place since 1975, and with the passing of the EHDI Act in 2007, all deaf and hard of hearing children have a birth stage in education.
Qwen2.5-VL:	I will show you how to do it.
Ground Truth:	Thank you!
SmolVLM2:	Congratulations!
Qwen2.5-VL:	Congrats and Congratulations!
Ground Truth:	I'm Crystal Cousineau
SmolVLM2:	I'm Crystal Cousineau
Qwen2.5-VL:	CRY.
Ground Truth:	The gunman is behind bars being held without bail
SmolVLM2:	It took about that same time for his owner to get down on one knee and say \"I love you\" in Spanish.
Qwen2.5-VL:	The suspect was arrested and is 17 years old
Ground Truth:	I'm Candace Jones.
SmolVLM2:	Thank you for watching Sign1News.
Qwen2.5-VL:	I'm Crystal Cousineau.

Table 11: Translation examples from test set of **OpenASL** dataset, where we highlight the partial matches in blue.

Ground Truth:	What is the RAT test?
SmolVLM2:	The cost of the surgery is around Rs 2 lakhs.
Qwen2.5-VL:	and now there is an update
Ground Truth:	Have a look at images.
SmolVLM2:	Have a look at images.
Qwen2.5-VL:	She is a 30 year old woman
Ground Truth:	The priest sanctified the church with a special ceremony.
SmolVLM2:	The priest sanctified the church with a special ceremony.
Qwen2.5-VL:	The priest sanctified the church with a
Ground Truth:	and the ingredients should only be added when the fire is burning properly
SmolVLM2:	so that you can get all your essential nutrients
Qwen2.5-VL:	ISH NEWS has already made a video on this, you can watch it on our YouTube channel.
Ground Truth:	People can not step out for frivolous matters.
SmolVLM2:	on flights to and from the UK said that,
Qwen2.5-VL:	People can not step out for frivolous matters.
Ground Truth:	the process of getting something
SmolVLM2:	the process of getting something that you want
Qwen2.5-VL:	acquisition noun

Table 12: Translation examples from test set of **iSign** dataset, where we highlight the partial matches in blue.

C Error Patterns in Noun Phrase

The qualitative examples in Tables 13 and 14 highlight a limitation in SmolVLM2's ability to reliably identify and preserve key noun phrases during translation. We present a few examples in the PHOENIX-2014T and OpenASL datasets, which show that the model often substituted critical entities with semantically plausible but incorrect alternatives (e.g., "rehabilitation act" → "education program", "Scandinavia" → "alps"). In some cases, like for "july" and "freitag", the entities are correctly identified, yet the model frequently misses domain-relevant constructs (e.g., legal terms, geographic regions) that are essential for meaningful interpretation. These patterns point to a broader learning challenge: SmolVLM2 generates fluent text but struggles to produce the correct entities in many cases. Gloss identification by the model should be improved to reduce such hallucinations.

Noun Phrase	PHOENIX-2014T
Ground Truth	"skandinavien", "wolken"
(English Translation)	("Scandinavia", "clouds")
Predicted (SmolVLM2)	"alpen", "norden", "nacht", "schnee"
(English Translation)	("alps", "north", "night", "snow")
Ground Truth	"westen", "sonne"
(English Translation)	("west", "sun")
Predicted (SmolVLM2)	"westen", "einzelne schauer", "nordwesten"
(English Translation)	("west", "isolated showers", "northwest")
Ground Truth	"neunundzwanzigsten oktober", "wettervorhersage"
(English Translation)	("twenty-ninth October", "weather forecast")
Predicted (SmolVLM2)	"fünften dezember", "wettervorhersage"
(English Translation)	("fifth of December", "weather forecast")
Ground Truth	"nordwesten deutschlands", "ostsee"
(English Translation)	("northwest Germany", "Baltic Sea")
Predicted (SmolVLM2)	"westen", "kurzen wetterberuhigung", "neuer tiefausläufer"
(English Translation)	("west", "brief weather calm", "new low pressure system")
Ground Truth	"norden", "alpen", "freitag", "süden", "einzelne schauer"
(English Translation)	("north", "alps", "friday", "south", "isolated showers")
Predicted (SmolVLM2)	"freitag", "einzelne schauer", "sonne", "längere zeit"
(English Translation)	("Friday", "isolated showers", "sun", "longer period")

Table 13: Comparison of extracted noun phrases from ground truth and SmolVLM2 predictions in the PHOENIX-2014T dataset. Matching entities are highlighted in blue.

Noun Phrase	OpenASL
Ground Truth	"section", "suit", "rehabilitation act", "title ii", "americans", "disabilities act"
Predicted (SmolVLM2)	"children", "education", "program", "nad", "projects", "child", "board members"
Ground Truth	"issues", "demands", "city"
Predicted (SmolVLM2)	"conference", "hartford", "ndec", "deaf education conference"
Ground Truth	"employers", "state laws", "employees"
Predicted (SmolVLM2)	"license", "driver", "people"
Ground Truth	"soil"
Predicted (SmolVLM2)	"everything", "earth", "flowers", "thing", "mindset", "trees", "nature", "plants"
Ground Truth	"july", "court", "testimony"
Predicted (SmolVLM2)	"july", "wednesday", "plan", "board"

Table 14: Examples from the OpenASL test set illustrating matches and mismatches between ground-truth and predicted noun phrases. Matched entities are highlighted in blue. SmolVLM2 rarely aligns correctly (e.g., "july"); frequently it hallucinates incorrect content.

D Vocabulary Bias and Distributional Influence

Tables 15 shows some of the most frequent noun phrases in the predicted text (for the test set), and their actual frequency in the ground-truth translations (for the same test set) and in the training set. We observe that many top-predicted noun phrases, such as *süden*, *norden*, and *tag*, have higher frequencies in the predicted set than in the ground-truth test set, and they also have relatively

higher frequency in the training set. Similarly, Table 16 shows the most frequent noun phrases in the ground-truth translations (in the test set), and their frequencies in the model's output, the ground-truth test set, and in the training set. We observe the same pattern that the more frequent words in the training data were over-represented in the model's output. This is a classic example of data imbalance, which makes model training very challenging. These biases in the dataset should be addressed so that model training can be done effectively.

Noun Phrase	Predicted Freq	Test Freq	Train Freq
es	191	110	1419
süden	106	50	653
norden	67	54	628
tag	62	31	422
uns	48	37	331
wettervorhersage	43	42	354
osten	41	30	418
sich	40	70	822
montag	32	17	194
westen	31	35	520

Table 15: Top 10 noun phrases from the predicted outputs with their frequencies in the predicted, test, and training sets.

Noun Phrase	Predicted Freq	Test Freq	Train Freq
es	191	110	1419
sich	40	70	822
norden	67	54	628
süden	106	50	653
wettervorhersage	43	42	354
uns	48	37	331
westen	31	35	520
tag	62	31	422
osten	41	30	418
samstag	28	28	242

Table 16: Top 10 noun phrases from the test set along with their frequency in the predicted and training sets.

Overall, these analyses indicate that training-induced vocabulary bias significantly influenced generation quality. The model often preferred frequent but potentially irrelevant tokens, leading to degraded semantic precision. These findings underline the need for frequency-aware modeling strategies, particularly for low-resource and long-tailed sign language datasets.

E Prompt Used

The prompts used to fine-tune both of our models, and also during inference, are present in the Table 17.

Dataset	Prompt
DGS (PHOENIX-2014T)	Übersetzen Sie das Video in deutscher Gebärdensprache in deutschen Text (gesprochene Sprache).
English Translation	Translate the German Sign Language video into German text (spoken language).
ASL (How2Sign and OpenASL)	Translate the English sign language video into spoken text.
ISL (iSign)	Translate the Indian sign language video into spoken text.

Table 17: The prompt used across different datasets, for fine-tuning and inference.

F Effect of Frame Length

Table 18 indicates that using more input frames yields only marginal gains across BLEU and ROUGE-L scores. This suggests that, although InternVL3 can benefit from additional temporal context, the improvements are not substantial enough to justify the significantly higher computational cost of longer video inputs.

Max Frames	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE-L
64	25.97	15.65	9.93	7.997	26.49
256	**26.56**	**16.13**	**10.71**	**8.79**	**26.91**

Table 18: Impact of Input Frame Length on Translation Performance for InternVL3 Fine-Tuned on PHOENIX-2014T.

G Computational Resources

We conducted our experiments on a workstation with Intel® Xeon® Processor ICX 6326 @ 2.9 GHz, 16 Cores, 32 Threads, 256.0 GB RAM, NVIDIA A100 GPU, CUDA Version: 12.6, and Ubuntu 22.04 operating system.