Seeing, Signing, and Saying: A Vision-Language Model-Assisted Pipeline for Sign Language Data Acquisition and Curation from Social Media

Shakib Yazdani[1] **Yasser Hamidullah**[1] **Cristina España-Bonet**[1,2] **Josef van Genabith**[1]
{shakib.yazdani,yasser.hamidullah,cristinae,josef.van_genabith}@dfki.de
[1]German Research Center for Artificial Intelligence (DFKI GmbH),
Saarland Informatics Campus, Saarbrücken, Germany
[2]Barcelona Supercomputing Center (BSC-CNS), Barcelona, Catalonia, Spain

Abstract

Most existing sign language translation (SLT) datasets are limited in scale, lack multilingual coverage, and are costly to curate due to their reliance on expert annotation and controlled recording setup. Recently, Vision Language Models (VLMs) have demonstrated strong capabilities as evaluators and real-time assistants. Despite these advancements, their potential remains untapped in the context of sign language dataset acquisition. To bridge this gap, we introduce the first automated annotation and filtering framework that utilizes VLMs to reduce reliance on manual effort while preserving data quality. Our method is applied to TikTok videos across eight sign languages and to the already curated YouTube-SL-25 dataset in German Sign Language for the purpose of additional evaluation. Our VLM-based pipeline includes a face visibility detection, a sign activity recognition, a text extraction from video content, and a judgment step to validate alignment between video and text, implementing generic filtering, annotation and validation steps. Using the resulting corpus, TikTok-SL-8, we assess the performance of two off-the-shelf SLT models on our filtered dataset for German and American Sign Languages, with the goal of establishing baselines and evaluating the robustness of recent models on automatically extracted, slightly noisy data. Our work enables scalable, weakly supervised pretraining for SLT and facilitates data acquisition from social media.[1]

1 Introduction

Sign languages are rich, expressive visual languages that serve as the primary means of communication for millions of Deaf and Hard-of-Hearing individuals worldwide. Sign languages convey meaning through a combination of hand configurations, movements, facial expressions, and body posture. Each sign language follows its own unique grammar and structure, making them distinct from one another (Stokoe, 1980).

Sign language translation (SLT), the task of translating a sign language video into spoken language, is evolving both in terms of dataset scale and the performance of SLT models. Despite the increasing availability of larger sign language datasets, such as YouTube-ASL (Uthus et al., 2023) for American Sign Language (ASL) and BOBSL (Albanie et al., 2021) for British Sign Language (BSL), as well as multilingual resources like JWSign (Gueuwou et al., 2023a) or SP-10 (Yin et al., 2022), the process of collecting, filtering, and annotating sign language data remains a labor-intensive and resource-demanding process. Given the requirement for proficient signers to ensure annotation accuracy, the scalability of such efforts is inherently limited. Social media focused efforts like YouTube-SL-25 (Tanzer and Zhang, 2025) have paved the way toward large-scale, open-domain multilingual SLT datasets, offering a partial solution to the challenges of scale and accessibility. Yet, the rapid development of foundation models—especially Vision-Language Models (VLMs)—presents an underexplored opportunity to further ease and scale the acquisition and annotation process by leveraging the vast amount of user-generated sign language content on platforms such as YouTube, TikTok, Pinterest, and Instagram.

Building on this direction, recent advancements in sign language translation have increasingly relied on self-supervised pre-training followed by task-specific fine-tuning. Studies such as (Hamidullah et al., 2024; Rust et al., 2024; Liang et al., 2024; Zhang et al., 2024; Li et al., 2025b) have demonstrated the effectiveness of this approach, enabling models to learn meaningful representations from

---

[1]The dataset (TikTok video IDs) and code for our pipeline is available at https://github.com/shakibyzn/vlm-sign-curator.

unlabeled data before being adapted to downstream tasks. Furthermore, even pretraining on weakly labeled data has been shown to enhance model performance, highlighting the potential of leveraging large-scale but noisier supervision signals (Rust et al., 2024).

To this end, we introduce a novel vision-language model-assisted pipeline to streamline data acquisition and filtering for SLT, with a particular focus on social media platforms such as Tik-Tok. As illustrated in Figure 1, the core idea behind our framework is straightforward: leveraging the distinction between manual (hand movements) and non-manual (facial expressions) components of sign language (Núñez-Marcos et al., 2023), and observing similar structural patterns in social media content (e.g., the presence of video descriptions, hashtags, comments, and sometimes captions), we design a three-stage pipeline. Our pipeline uses the Qwen2.5-VL model to process videos in three stages: (1) it checks whether the subject's face is clearly visible (**FaceDetector**) and whether the person exhibits signs of signing activity (**SignActivityDetector**); (2) it extracts visible on-screen text, which often corresponds to the creator's intended translation (**TextExtractor**); and (3) it evaluates the alignment between the extracted text and the signing using Phi-4-Multimodal (Microsoft et al., 2025) as a model-as-a-judge (**Judge**). In all, our contributions are as follows:

- We present the first VLM-assisted pipeline for SLT data curation from social media, with a focus on TikTok. Our method focuses on capturing the core elements of sign language: face visibility, signing activity, and the corresponding spoken language translation.

- We release the TikTok-SL-8 dataset, comprising approximately 49 hours of video across 8 sign languages, curated through an automated VLM-assisted pipeline. Our method achieves an accuracy of 0.75 on DGS and 0.82 on ASL, closely matching human annotator performance.

- We additionally evaluate our dataset on DGS and ASL using two off-the-shelf SLT models to assess its performance.

2 Related Work

In this section, we review existing SLT datasets and methodologies for translating sign languages into spoken language text. Additionally, we highlight the emerging role of LLM-assisted techniques in data curation and filtering.

2.1 Sign Language Translation Datasets

One of the most widely used sign language translation datasets is RWTH-PHOENIX-2014T (Camgoz et al., 2018), which contains 11 hours of German Sign Language (DGS) weather forecasts interpreted from the German TV station PHOENIX. In recent years, more TV broadcast datasets have emerged, expanding the scope of sign language research. SWISSTXT (Camgoz et al., 2021) provides 152 hours of news and weather programs interpreted into Swiss German Sign Language, while the BOBSL (Albanie et al., 2021) dataset contains 1,447 hours of British Sign Language (BSL) interpretations from various BBC programs across multiple domains. These datasets play a crucial role in advancing sign language translation by providing large-scale, real-world training resources. Other datasets involve signers translating predefined phrases in controlled environments or using personal recording devices. CSL-Daily (Zhou et al., 2021), comprising 23 hours of content, focuses on daily life phrases in Chinese Sign Language (CSL), while How2Sign (Duarte et al., 2021), with over 80 hours of data, provides instructional monologues in ASL. Moreover, multilingual sign language datasets are emerging, though many remain limited in scope and domain. AfriSign (Gueuwou et al., 2023b) and JWSign (Gueuwou et al., 2023a) focus on Bible translations, while SP-10 (Yin et al., 2022) includes 10 sign languages but primarily features very short sentences. In contrast, YouTube-SL-25 (Tanzer and Zhang, 2025) is a large-scale, open-domain dataset with over 3,000 hours of content across more than 25 sign languages.

2.2 Sign Language Translation Models

Traditional SLT models have relied on gloss annotations as an intermediate representation, but recent advances have explored end-to-end models, self-supervised learning, and the integration of large language models (LLMs).

Camgoz et al. (2018) introduced the task of SLT as an end-to-end solution that jointly incorporated glosses and evaluated their method on the PHOENIX14T dataset. They later extended it to a gloss-free model (Camgoz et al., 2020) that directly mapped sign videos to spoken sentences. As gloss annotation is labor-intensive, research has shifted towards gloss-free SLT. Tarrés et al. (2023) fur-

thered this by using I3D visual features (Carreira and Zisserman, 2017) to train an SLT model on the How2Sign dataset. Additionally, some works have framed SLT as a self-supervised learning problem, including sentence-level embedding-based supervision (Hamidullah et al., 2024), scaling datasets and models (Zhang et al., 2024), and addressing data scarcity concerns in large-scale privacy-aware SLT training (Rust et al., 2024). Radford et al. (2021) leveraged visual-language pretraining by integrating masked self-supervised learning with CLIP to enable gloss-free SLT. Recent SLT research has increasingly incorporated advancements in LLMs and multimodal learning. The SignLLM framework (Gong et al., 2024) transforms sign videos into a structured representation for improved processing by LLMs. LLaVA-SLT (Liang et al., 2024) employs large multimodal models by first performing linguistic continued pre-training on a sign language corpus to refine linguistic capabilities. This is followed by contrastive learning to align visual and textual representations, and finally, integration via a lightweight multi-layer perceptron connector for SLT. While prior work has focused on using LLMs for translation, to the best of our knowledge, we are the first to leverage LLM-assisted methods for data collection and annotation in the SLT domain.

2.3 LLM-assisted Data Collection

Collecting and annotating large-scale multilingual sign language datasets remains a significant challenge due to the complexity, subjectivity, and resource-intensive nature of the process. Traditional methods rely heavily on human annotators with domain expertise, making large-scale data curation costly and time-consuming. Recent advancements in multimodal large language models (MLLMs), such as Qwen-2.5-VL (Bai et al., 2025), Phi-4-multimodal (Microsoft et al., 2025) and DeepSeek-VL2 (Wu et al., 2024), present a novel opportunity to automate this process by either supplementing or fully replacing human annotators with the goal of minimizing reliance on manual labor. While such approaches have been successfully explored in other domains, they remain untapped in the context of SLT data collection and curation. While not situated in the SLT domain, several recent works have explored the use of LLMs to assist with multimodal data collection and curation. For example, Choi et al. (2024) introduces the

VOLDOGER dataset, which leverages multimodal LLM-based annotation to streamline data labeling for vision-language tasks like image captioning, visual question answering (VQA), and visual entailment (VE), reducing reliance on human annotators. Wang et al. (2024) introduces the Model-in-the-Loop (MILO) framework, which integrates LLMs into the data annotation process to improve efficiency and quality by assisting human annotators. They experiment with a multimodal dataset of social media comments and demonstrated that MILO reduced annotation time and enhanced labeling efficiency. Yeh et al. (2024) collected the COCOLOFA dataset with the assistance of crowd workers and LLMs.
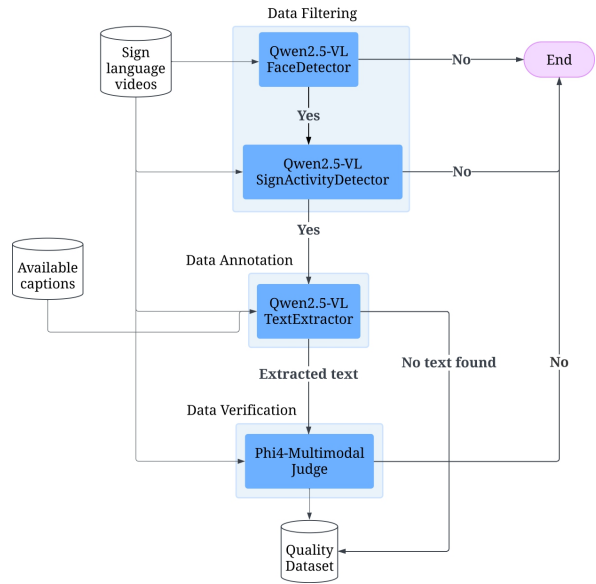

Figure 1: An overview of our VLM-based SLT dataset collection framework on social media, with a particular focus on TikTok. The pipeline consists of three key stages: data filtering, data annotation, and data verification.

3 Method

In this section, we introduce our framework, which leverages the multimodal large language model `Qwen-2.5-VL` for **filtering** and **annotation** and `Phi-4-multimodal` as a **model-as-a-judge**, as illustrated in Figure 1. After initial data collection, our framework consists of three key stages: Data Filtering, Annotation, and Verification. In the following sections, we will explore each stage in detail.

3.1 Data Collection

As motivated in the introduction, our approach adopts an social media data source setting similar to YouTube-SL-25, but extends it to a more challenging platform—TikTok—which contains potentially noisier content. Unlike YouTube, which primarily features longer, educational content, TikTok is characterized by short-form, trend-driven videos. This shift introduces additional challenges, including inconsistent caption quality and greater variability in content styles.

Our goal is to develop a generalizable framework for SLT data acquisition and filtering. Based on our observation that social media videos tend to follow similar structural patterns—such as the presence of video descriptions, hashtags, comments, and occasionally captions—we implement two starting point strategies for automatic data collection:

1. **Hashtag-based retrieval**,[2] where we identify relevant hashtags.

2. **User-based retrieval**, where we identify high-quality users who consistently produce sign language content.

We manually select hashtags by inspecting TikTok video descriptions, ensuring that each hashtag appears in both English and the native language associated with the target sign language (e.g., for DGS, hashtags include both "Germansignlanguage" and "Gebärdensprache"). High-quality users are identified through manual inspection of profiles in TikTok's top recommendations. Due to privacy concerns, we do not disclose the usernames of these individuals.

3.2 Data Filtering

We design and rely on well-crafted prompts for VLM-based filtering in order to automate and scale the identification of sign language unrelated content. To this end, we use `Qwen-2.5-VL` as the VLM-based **FaceDetector** and **SignActivityDetector**, due to its strong performance on video understanding tasks (Bai et al., 2025). The VLM **FaceDetector** analyzes the video frames to determine if a person (or multiple people) is present and whether their face is clearly visible and identifiable. If the face is obscured or unrecognizable, the video will be discarded. The specific prompt we used for the

FaceDetector stage is shown in Figure 2. The VLM **SignActivityDetector**, on the other hand, assesses whether the person in the video is primarily using sign language for communication, classifying individuals based on their signing behavior. [3] Together, these two stages help filter videos to ensure relevant and high-quality sign language content is selected.
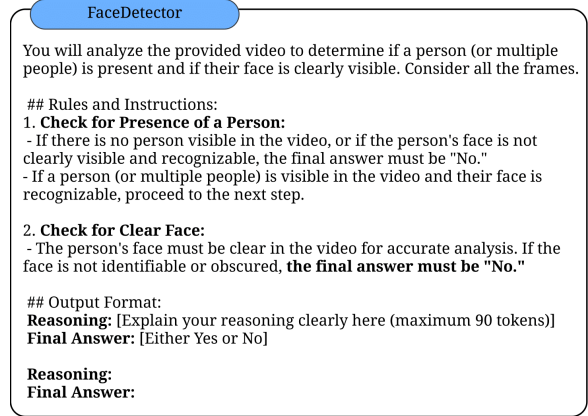
Figure 2: Prompt template used for the **VLM FaceDetector**.

3.3 Data Annotation

Unlike other datasets discussed in the related work section—such as those sourced from YouTube, which often come with seemingly well-aligned captions, or those collected in controlled lab environments with invited signers, resulting in high-quality but small and less diverse datasets—TikTok presents a unique set of challenges for sign language translation data curation. TikTok videos typically lack formal captioning. Instead, creators often embed text directly within the video itself. However, this text is not always a faithful translation of the signed content; it may simply include the title of a lesson, a promotional link, or the name of a song being signed. This ambiguity complicates the task of aligning signed content with corresponding text. To address these challenges while benefiting from TikTok's diversity and real-world context, we again use `Qwen2.5-VL` for its video optical character recognition (OCR) and understanding capabilities. Specifically, we use a vision-language module called **TextExtractor**, which identifies and extracts text from videos featuring sign language. It verifies the extracted text for readability and alignment with the video content, prefers formal captions when

---

[2]The full list of hashtags is provided in the GitHub repository.

[3]We provide the prompt used for SignActivityDetector stage in our GitHub repository.

available, and flags instances where no textual content is found. [4]

3.4 Data Verification

To ensure the accuracy and relevance of our framework, we adopt a "model-as-a-judge" approach using a VLM **Judge** as the final verification step. As previously discussed, while some TikTok videos include captions, these captions may not accurately reflect the signed content. In many cases, the text embedded within the video may be unrelated—serving instead as a lesson title, app promotion, or song title—rather than a true translation of the signs. To address this issue, we employ `Phi-4-Multimodal` as a VLM **Judge**. In accordance with best practices in "model-as-a-judge" approaches, and to mitigate common problems such as *self-preference bias* (Wataoka et al., 2024) and *preference leakage* (Li et al., 2025a), we intentionally use a model different from `Qwen2.5-VL` for the evaluation stage. `Phi-4-multimodal` has also demonstrated stronger performance than `Qwen2.5-VL` on both the multi-image benchmark BLINK (Fu et al., 2024b) and the video benchmark VIDEOMME (Fu et al., 2024a), making it well-suited to analyze the signed content, determine its meaning, and verify whether it matches the provided caption—ensuring the resulting dataset remains both accurate and reliable. [5]

3.5 Resulting Dataset

As highlighted in the introduction, in terms of content we adopt an social media setting similar to YouTube-SL-25. Our goal in building **TikTok-SL-8** is to reduce the manual effort typically required from human annotators by leveraging the capabilities of multimodal LLMs. The final dataset, constructed using our fully automatic three-stage framework across eight sign languages, is summarized in Table 1, which presents the total number of videos and hours of content per sign language. In total, TikTok-SL-8 comprises **49 hours** of sign language video content across 8 languages sourced from TikTok—surpassing the SP-10 dataset, which contains 14 hours across ten sign languages. Notably, the entire dataset was curated automatically using our pipeline. The only manual interventions involved were in selecting relevant hashtags, iden-

tifying high-quality content creators (see Section 3.1), and crafting prompt templates tailored to each stage of the framework.

Sign Language	ISO 639	#Videos	#Hours
American	ase	816	8
Australian	asf	541	4
British	bfi	1201	15
Chinese	csl	218	2
French	fsl	285	3
German	gsg	499	4
Italian	ise	678	7
Swedish	swl	562	6
Total	—	4800	49

Table 1: Statistics for different sign languages in the curated dataset, including ISO 639 codes, total number of videos, and total duration.

4 Experiments

We demonstrate the value of our VLM-based data acquisition and curation framework, as well as the quality of the resulting dataset for SLT, across two sign languages. Additionally, we describe how our automatic data curation framework was applied to YouTube-SL-25 (DGS) and evaluate its performance using a manually curated subset of the data as gold-standard annotations.

4.1 Implementation Details

For `Qwen2.5-VL`, we use the 7B variant, and for `Phi-4-multimodal`, we employ the `Phi-4-multimodal-instruct` model with 5.57B parameters available on Hugging Face. To process the videos, we adopt the default smart resizing method[6] provided by the `Qwen2.5-VL` series, which supports dynamic resolution and frame rate. In our setup, however, we fix the input resolution to a width and height of 224 pixels.

4.2 Data Acquisition and Curation Evaluation

To validate the quality of our framework, the first author[7] manually annotated approximately 150

---

[4]We provide the prompt template used for the VLM TextExtractor in our GitHub repository.

[5]We provide the prompt template employed for the VLM Judge in our GitHub repository.

[6]https://github.com/QwenLM/Qwen2.5-VL/blob/main/qwen-vl-utils/src/qwen_vl_utils/vision_process.py

[7]The first author is a non-native hearing annotator. Annotations were performed by comparing the signs in the video with any captions provided within the video, the automatic captions (if available), and user comments to assess consistency and semantic alignment.

samples from the DGS and ASL TikTok-SL-8 subsets included in the final dataset. We use Accuracy, Precision, and Recall to evaluate the overall performance of our framework in comparison to manual annotation. Furthermore, to demonstrate the applicability of our approach to other open-domain settings such as YouTube, we further evaluate the framework by applying it to the already curated YouTube-SL-25 (DGS). Table 2 presents the performance of our VLM-based framework across different datasets and sign languages. For the YouTube-SL-25 (DGS) gold dataset, we treat all samples as positive ground-truth labels, which makes precision not applicable. On the gold DGS data set, our framework achieves an accuracy and recall of 0.86. On our TikTok-SL-8 dataset, the framework shows solid performance across both DGS and ASL, with particularly high precision on ASL (0.91) and strong recall on DGS (0.87), indicating its effectiveness in diverse and noisy open-domain scenarios.

Dataset	Accuracy	Precision	Recall
YouTube (DGS)	0.86	–	0.86
TikTok (DGS)	0.75	0.72	0.87
TikTok (ASL)	0.82	0.91	0.79

Table 2: Performance of our framework across TikTok-SL-8 (TikTok) and YouTube-SL-25 (YouTube). Precision is not applicable to the YouTube dataset due to the absence of negative ground-truth labels.

Figure 3 presents the confusion matrices between our framework's predicted labels and manual annotations on the DGS and ASL subsets of the TikTok-SL-8 dataset. The confusion matrices show strong agreement, particularly for ASL, where the framework achieves 75 true positives and 50 true negatives, indicating high precision and recall. In DGS, the framework also performs well, with 71 true positives, though with a slightly higher false positive count.

4.3 SLT Experiments Using the Data

Baselines. We demonstrate the effectiveness of our VLM-based data acquisition and curation framework and the TikTok-SL-8 dataset using two gloss-free open-source baselines across two sign languages: SEM-SLT (Hamidullah et al., 2024) and Signformer (Yang, 2024).
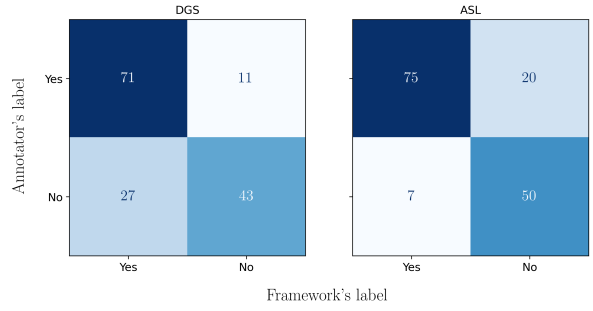

Figure 3: Confusion matrices showing the agreement between our framework's automated labels and manual annotations for DGS and ASL subsets of TikTok-SL-8.

SEM-SLT (Hamidullah et al., 2024). SEM-SLT consists of a visual encoder that maps videos to sentence embeddings (sign2sem) and a decoder that maps sentence embeddings to text (sem2text). Both components are pretrained separately and then combined and fine-tuned in an end-to-end fashion. We train the sign2em visual encoder to map sign language videos into sentence embeddings using video features extracted with EF-Net-B0 (Tan and Le, 2019) from the TikTok-SL-8 dataset. The sentence embeddings are intended to represent the overall meaning of the signed utterances.

The original sem2text module, which maps semantic embeddings to spoken language sentences, was designed for single-sentence outputs. However, since TikTok-SL-8 contains multi-sentence segments, we adapted the sem2text decoder accordingly. We first pre-trained the embedding projection layer and the first six layers of the mBART (Liu et al., 2020) decoder on the CNN/Daily Mail and MLSUM[de] datasets (See et al., 2017; Scialom et al., 2020). This pretraining step allowed the model to better handle longer text sequences. We then fine-tuned the mBART decoder on the target transcriptions from TikTok-SL-8 training data, aligning it with the TikTok-specific language and structure. Finally, we trained the full SEM-SLT pipeline in an end-to-end manner. This involved combining the visual encoder and the sem2text decoder into a single model—Sign2(sem+text)—and supervising it directly using the target TikTok-SL-8 training data sentence embeddings. This step refines the entire system to translate raw sign language video into spoken language text.

Signformer (Yang, 2024). Signformer is an encoder–decoder Transformer-based model designed for edge AI applications, achieving strong

performance with a compact size of only 0.57 million parameters. Like Hamidullah et al. (2024) SEM-SLT, Signformer is a gloss-free model that does not rely on any pre-training. Instead, it leverages a novel attention mechanism proposed by Yin et al. (2023), combined with a convolutional module. We train Signformer from scratch using features extracted by the S3D model (Xie et al., 2018), pre-trained on both the WLASL (Li et al., 2020) and Kinetics-400 (Kay et al., 2017) datasets. For feature extraction, we use only the first four blocks of the S3D architecture. Each input video is passed through the S3D encoder, and the output from the final block is spatially pooled to obtain a feature representation of size $F/4 \times 832$, where F denotes the number of frames in the video.

Evaluation Metrics. We evaluate translation performance using standard metrics: chrF (Popović, 2015), BLEU (Papineni et al., 2002), and BLEURT (Sellam et al., 2020). For BLEU[8] and chrF,[9] we use sacreBLEU (Post, 2018). For BLEURT, we use the official Python library.[10]

4.4 Results on TikTok-SL-8

To demonstrate the quality of our VLM-based data collection and curation framework, we evaluate on held out test sets from the ASL and DGS subsets of our resulting dataset, TikTok-SL-8. We consider only those videos that either include captions or for which our framework successfully extracted text. After this filtering step, the ASL subset contains 649 training and 75 test videos, while the DGS subset contains 391 training and 71 test videos. We report results using Signformer and SEM-SLT using the BLEURT, chrF, and BLEU metrics. Table 3 presents the quantitative results. The results show that SEM-SLT, a model incorporating pre-training, consistently outperforms Signformer, even when trained on noisy data that includes misaligned captions. While BLEURT and chrF scores are relatively close for both models, BLEU scores indicate a clear advantage for SEM-SLT. Overall, the findings show that, despite being collected through an automatic, VLM-assisted pipeline and containing noisy supervision, the TikTok-SL-8 dataset retains sufficient alignment between sign and text to

support effective training of sign language translation models. Overall, results indicate that the pretraining-based SEM-SLT model is better able to handle slightly noisy data than Signformer.

Dataset	Method	BLEURT	chrF	BLEU
TikTok (DGS)	Signformer	0.18 ± 0.04	16.1 ± 6.7	5.2 ± 4.9
	SEM-SLT	0.21 ± 0.04	18.8 ± 5.6	9.2 ± 5.9
TikTok (ASL)	Signformer	0.30 ± 0.03	13.1 ± 2.5	0.9 ± 0.7
	SEM-SLT	0.31 ± 0.02	15.8 ± 3.6	5.6 ± 4.3

Table 3: Performance of Signformer and SEM-SLT on the DGS and ASL subsets of TikTok-SL-8 (TikTok) using BLEURT, chrF, and BLEU.

Lang.	Text (German / English)
DGS	**Ref:** Hallo! Du hast bis jetzt nur Lautsprache gelernt. Französisch, Spanisch und so? Und jetzt noch mal eine Lautsprache lernen?. Nein, jetzt ist Gebärdensprache dran! *(Hello! Up until now, you've only learned spoken languages. French, Spanish, and so on? And now you're going to learn another spoken language? No, now it's time for sign language!)* **Signformer:** ich liebe dich. *(I love you.)* **SEM-SLT:** Hallo! Du hast bis jetzt nur Lautsprache gelernt. Französisch, Spanisch und so? Und jetzt noch mal eine Lautsprache lernen? Nein, jetzt ist Gebärdensprache dran! *(Hello! Up until now, you've only learned spoken languages. French, Spanish, and so on? And now you're going to learn another spoken language? No, now it's time for sign language!)*
	Ref: Mein Name Mein Name Alles gut. Alles gut. gut gut Dein Name was? Dein Name was? *(My name, my name. All good. All good. Good, good. Your name what? Your name what?)* **Signformer:** alles gut? gut. dein name was? mein name. *(All good? Good. Your name what? My name.)* **SEM-SLT:** Ich bin froh, dass du hier bist. *(I am glad that you are here.)*
ASL	**Ref:** The sign of the day is: try, attempt, effort. Try. For example, "If, at first, you don't succeed. Try, try again." **Signformer:** the sign of the day is ongoing or continue. on go. ongoing. for example, "you want to become proficient at american sign language? well, i'll tell you how to do it socialize with deaf individuals, pick up on the signs that they use, and accept their feedback. it's an ongoing process." **SEM-SLT:** ASL is more. Try. For example, maybe more. Try. For example, at the end of the day you're gonna try to try again. Try my body is shorten to the day. Try. It'
	Ref: Today I learned School, School. **Signformer:** cc today i learned school, school. **SEM-SLT:** Today I Learned school. School

Table 4: Example translations from TikTok-SL-8 for DGS and ASL subsets using Signformer and SEM-SLT models.

4.5 Qualitative Analysis

Table 4 presents two example translations each from the DGS and ASL subsets of our TikTok-SL-8 dataset, using the Signformer and SEM-SLT models. In the DGS examples, SEM-SLT slightly outperforms Signformer in translation quality, aligning with the quantitative results shown in Table 3. In the ASL examples, SEM-SLT again demonstrates more accurate translations. This aligns with the

---

[8]BLEU|nrefs:1|case:mixed|eff:no|tok:13a| smooth:exp|version:2.5.1

[9]chrF|nrefs:1|case:mixed|eff:yes|nc:6|nw:0| space:no|version:2.5.1

[10]BLEURT using checkpoint BLEURT-20.

BLEU score of 5.59 reported for ASL in Table 3. Manual analysis confirms that the pretraing-based SEM-SLT is better able to use slightly noisy data than Signformer.

4.6 Ablation Study

To further assess the effectiveness of our approach, we conducted an additional ablation study on both our data collection and curation framework, and the resulting dataset.

Caption Agreement. Although `Qwen2.5-VL-7B` demonstrated promising OCR capabilities, we further evaluated its performance on the DGS and ASL subsets of TikTok-SL-8 by measuring caption agreement between our framework (VLM TextExtractor) and human annotator translations. We used BLEURT, chrF, and BLEU as evaluation metrics. Videos for which our framework returned "No text found." were excluded from the analysis. After this filtering step, the held-out test sets contained 65 samples for the ASL subset and 66 for the DGS subset. Table 5 shows a high degree of agreement across both subsets, with notably higher alignment in the ASL subset.

Dataset	BLEURT	chrF	BLEU
TikTok (DGS)	0.53 ± 0.08	66.5 ± 13.3	52.1 ± 17.5
TikTok (ASL)	0.71 ± 0.06	83.6 ± 9.3	77.0 ± 13.6

Table 5: Caption agreement between framework-extracted captions and human annotator translations on the DGS and ASL subsets of TikTok-SL-8, measured using BLEURT, chrF, and BLEU.

5 Conclusion

In this paper, we introduced the first automated data acquisition and curation framework based on Vision-Language Models for sign language translation from social media. Our approach simplifies dataset collection for sign language translation and supports human annotators by implementing a three-stage pipeline. Our pipeline leverages `Qwen2.5-VL` as a FaceDetector, SignActivityDetector, and TextExtractor, and employs `Phi-4-multimodal` as a *Model-as-a-Judge* to assess whether the extracted text aligns with the signing in the video. We evaluated our VLM-based data acquisition and curation framework through two complementary approaches. First, we assessed the quality of automatically extracted data by comparing it against manually annotated subsets of our

TikTok-SL-8 dataset for both ASL and DGS, as well as demonstrating how our framework performs on the already curated YouTube-SL-25 gold dataset for DGS. This evaluation allows us to benchmark the effectiveness of our VLM-based pipeline. Second, we evaluated the utility of the curated data by training two open-source gloss-free SLT models—*Signformer* and *SEM-SLT*. Results indicate that the pretraining-based SEM-SLT model is better able to handle slightly noisy data than Signformer. Furthermore, we demonstrated that the captions extracted by our framework exhibit a high degree of agreement with human-translated references. We believe our VLM-based framework can streamline the acquisition of large amounts of sign language data already available on social media platforms.

Limitations

While our automated data acquisition and curation framework for sign language translation from social media shows promising results, there are both limitations and opportunities for improvement. In particular, we rely on VLMs for two critical stages: sign activity recognition and text extraction. Despite notable progress in OCR tasks with VLMs, significant challenges remain—particularly in handling longer videos due to context length limitations. These limitations can lead to misclassifications, such as incorrectly labeling hearing individuals as signers, or generating captions that are poorly aligned with the actual content. Additionally, our approach primarily relied on prompt engineering, which may not yield optimal results. More sophisticated prompting strategies, such as Chain-of-Thought prompting (Wei et al., 2022), could potentially enhance our framework.

References

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. 2021. BOBSL: BBC-Oxford British Sign Language Dataset. *arXiv*.

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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

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)*, pages 10023–10033.

Necati Cihan Camgoz, Ben Saunders, Guillaume Rochette, Marco Giovanelli, Giacomo Inches, Robin Nachtrab-Ribback, and Richard Bowden. 2021. Content4All Open Research Sign Language Translation Datasets . In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–5, Los Alamitos, CA, USA. IEEE Computer Society.

João Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733.

Juhwan Choi, Junehyoung Kwon, JungMin Yun, Seunguk Yu, and YoungBin Kim. 2024. Voldoger: Llm-assisted datasets for domain generalization in vision-language tasks. *ArXiv*.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2735–2744.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2024a. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024b. Blink: Multi-modal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer.

Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. 2024. Llms are good sign language translators. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18362–18372.

Shester Gueuwou, Sophie Siake, Colin Leong, and Mathias Müller. 2023a. JWSign: A highly multilingual corpus of Bible translations for more diversity in sign language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9907–9927, Singapore. Association for Computational Linguistics.

Shester Gueuwou, Kate Takyi, Mathias Müller, Marco Stanley Nyarko, Richard Adade, and Rose-Mary Owusuaa Mensah Gyening. 2023b. Afrisign: Machine translation for african sign languages. In *4th Workshop on African Natural Language Processing*.

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

Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Apostol Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The kinetics human action video dataset. *ArXiv*, abs/1705.06950.

Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. 2025a. Preference leakage: A contamination problem in llm-as-a-judge. *arXiv preprint arXiv:2502.01534*.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1459–1469.

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

Han Liang, Chengyu Huang, Yuecheng Xu, Cheng Tang, Weicai Ye, Juze Zhang, Xin Chen, Jingyi Yu, and Lan Xu. 2024. LLaVA-SLT: Visual Language Tuning for Sign Language Translation. *ArXiv*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. 2025. Phi-4-mini technical report: Compact yet powerful multi-modal language models via mixture-of-loras. *ArXiv*.

Adrián Núñez-Marcos, Olatz Perez de Viñaspre, and Gorka Labaka. 2023. A survey on sign language machine translation. *Expert Systems with Applications*, 213:118993.

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. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395. Association for Computational Linguistics.

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.

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 *International Conference on Machine Learning*, pages 8748–8763. PmLR.

Phillip Rust, Bowen Shi, Skyler Wang, Necati Cihan Camgoz, and Jean Maillard. 2024. Towards privacy-aware sign language translation at scale. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8624–8641, Bangkok, Thailand. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, et al. 2020. Mlsum: The multilingual summarization corpus. In *2020 Conference on Empirical Methods in Natural Language Processing: Proceedings of the Conference*, pages 8051–8067. Association for Computational Linguistics (ACL).

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

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.

William C. Stokoe. 1980. Sign language structure. *Annual Review of Anthropology*, 9:365–390.

Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.

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

Laia Tarrés, Gerard I. Gállego, Amanda Duarte, Jordi Torres, and Xavier Giró i Nieto. 2023. Sign language translation from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): Workshops*, pages 5625–5635.

David Uthus, Garrett Tanzer, and Manfred Georg. 2023. Youtube-asl: a large-scale, open-domain american sign language-english parallel corpus. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Yifan Wang, David Stevens, Pranay Shah, Wenwen Jiang, Miao Liu, Xu Chen, Robert Kuo, Na Li, Boying Gong, Daniel Lee, Jiabo Hu, Ning Zhang, and Bob Kamma. 2024. Model-in-the-loop (milo): Accelerating multimodal ai data annotation with llms. *ArXiv*.

1383

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *ArXiv*.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 305–321.

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

Min-Hsuan Yeh, Ruyuan Wan, and Ting-Hao Kenneth Huang. 2024. CoCoLoFa: A dataset of news comments with common logical fallacies written by LLM-assisted crowds. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 660–677, Miami, Florida, USA. Association for Computational Linguistics.

Aoxiong Yin, Zhou Zhao, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. 2022. Mlslt: Towards multilingual sign language translation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5099–5109.

Aoxiong Yin, Tianyun Zhong, Li Tang, Weike Jin, Tao Jin, and Zhou Zhao. 2023. Gloss Attention for Gloss-free Sign Language Translation . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2551–2562, Los Alamitos, CA, USA. IEEE Computer Society.

Biao Zhang, Garrett Tanzer, and Orhan Firat. 2024. Scaling sign language translation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Hao Zhou, Wen gang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1325.