TRIVIS: A LARGE-SCALE MULTI-VIEW BENCHMARK FOR VIETNAMESE SIGN LANGUAGE RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper presents TriViS, the first large-scale multi-view Vietnamese dataset for Continuous Sign Language Recognition and Translation, designed as a controlled benchmark for studying dataset design in underrepresented sign languages. TriViS contains 12,000 sentences and 84,453 synchronized videos captured from three viewpoints (frontal, 45° left, and 45° right) across both controlled studio recordings and outdoor environments. The synchronized multi view setup enables systematic analysis of viewpoint sensitivity, view complementarity, and other data collection factors that are difficult to isolate in uncontrolled web-scale corpora. Along with the dataset, we introduce a simple multi-view fusion framework that integrates synchronized visual streams into a unified representation for sentence-level recognition and translation. Using state-of-the-art CSLR and CSLT baselines, we conduct extensive experiments to evaluate the contribution of individual viewpoints, multi-view fusion strategies, sentence-independent generalization, and the effect of LLM scale. Our results show that multi-view observations consistently outperform single-view baselines and provide new empirical insights into how viewpoint configuration and controlled data collection affect continuous sign language understanding. Our code and dataset are publicly available on GitHub and Hugging Face.

1 INTRODUCTION

Background. For the Deaf and hard-of-hearing community, sign language is far more than a collection of gestures; it is a fully developed natural language with its own complex phonology, morphology, and syntax (Stokoe Jr, 2005; Klima & Bellugi, 1979; Sandler et al., 2006). Automatic Continuous Sign Language Recognition (CSLR) and Translation (CSLT) systems therefore hold enormous potential as assistive technologies, enabling applications such as real-time classroom transcription, accessible healthcare, and seamless digital workplace integration. However, despite rapid progress in vision-language modeling, current CSLR systems remain far from the reliability required for everyday deployment (Ibrahim et al., 2020). Our preliminary evaluations in Table 3 using SOTA VLMs reveal that CSLR accuracy remains strikingly low: even strong vision-language backbones such as Qwen3-VL (Bai et al., 2025) and LLaVA-Video-7B (Zhang et al., 2024) achieve only 3.62 BLEU-4 on the laboratory subset and 2.03 BLEU-4 on the real-world subset.

One major reason for this performance plateau is that robust sign language understanding depends critically on the availability of suitable large-scale datasets. This challenge is significantly magnified when extending CSLR beyond a small number of well-studied languages like American (ASL) (Shi et al., 2022), British (BSL) (Albanie et al., 2021), or Chinese Sign Language (CSL) (Zhou et al., 2021). Our experiments (Section 5.1, Table 2) demonstrate a stark performance gap between high-resource and low-resource sign languages; for instance, GPT-5 (Singh et al., 2025) obtains 3.10 BLEU-4 on How2Sign (Duarte et al., 2021), an American Sign Language dataset, but decreases to 2.44 on TriViS, our Vietnamese Sign Language dataset, while Gemini-2.5 (Comanici et al., 2025) shows an even sharper decline from 2.70 to 1.40 BLEU-4. Consequently, bridging this resource gap is not just a technical necessity but a matter of global digital equity.

Developing CSLR systems for underrepresented languages is not simply a matter of transferring existing models; it requires a fundamental understanding of how to construct datasets under realistic collection constraints. Before investing in massive data collection, it is vital to identify which dataset properties, such as viewpoint and diversity, most strongly influence performance to prioritize data collection efforts effectively.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

CSLR Datasets and Limitations. Collecting large-scale continuous sign language datasets is inherently difficult. Unlike isolated sign recognition, CSLR requires long continuous signing sequences with accurate sentence-level annotations, synchronized temporal alignment, and sufficient signer diversity. Existing video-based CSLR datasets are mainly collected through two strategies. The first and most common strategy is to harvest sign language videos from broadcast television or web repositories, as in BOBSL (Albanie et al., 2021), LSA-T (Dal Bianco et al., 2022), and OpenASL (Shi et al., 2022). This approach enables large-scale data collection, but often introduces uncontrolled factors such as inconsistent video quality, cluttered backgrounds, varying viewpoints, occlusions, and limited visibility of fine-grained hand and facial motions. More importantly, these factors are tightly entangled, making it difficult to systematically study which recording conditions and visual signals are most important for robust recognition. The second strategy is data collection in controlled recording environments, as in How2Sign (Duarte et al., 2021) and CSL-Daily (Zhou et al., 2021). Although this approach is more expensive and labor-intensive, it provides clearer visual signals, synchronized recordings, and consistent acquisition conditions that enable controlled analysis of dataset design factors.

Despite these efforts, existing datasets suffer from two critical flaws. First, they exhibit a severe linguistic bias toward a few dominant languages such as American Sign Language (ASL) (Shi et al., 2022; Duarte et al., 2021), British Sign Language (BSL) (Albanie et al., 2021), Auslan (Shen et al., 2023), and Chinese Sign Language (CSL) (Zhou et al., 2021), leaving low-resource languages such as Vietnamese Sign Language substantially underrepresented and contributing to the poor performance mentioned above. Second, they suffer from a "frontal-view bias": most datasets are captured primarily from a single frontal perspective, limiting their ability to model signs whose discriminative cues depend on depth, hand orientation, or side-view motion dynamics.

Our TriViS Dataset. To bridge these gaps, we introduce TriViS, the first large-scale Vietnamese multi-view dataset for Continuous Sign Language Translation. We position VSL as a representative low-resource setting to study how future CSLR datasets can be constructed for underrepresented languages more broadly. Rather than focusing solely on maximizing accuracy, TriViS is designed as a controlled research platform to investigate viewpoint sensitivity and multi-view complementarity. Although smaller than "web-scale" corpora, its synchronized multi-view setting and controlled acquisition allow for systematic analysis that is impossible on uncontrolled datasets.

To support these investigations, we develop a simple yet effective multi-view framework that integrates synchronized left, frontal, and right visual streams into a unified representation. Using this framework, we conduct extensive experiments with SOTA models to analyze the benefits of multi-view data and the influence of LLM scale. Our goal is to advance Vietnamese CSLR while providing a blueprint for constructing deployment-relevant datasets for the world's underrepresented sign languages.

Our Contributions. The main contributions of this work are threefold:

- First, we introduce TriViS, the first large-scale multi-view Vietnamese dataset for CSLR/CSLT, featuring $12,000$ sentences and $84,453$ synchronized videos captured from frontal, 45° left, and 45° right viewpoints across 14,133 glosses.
- Second, we establish a synchronized multi-view collection framework that enables systematic analysis of viewpoint sensitivity in continuous sign language recognition. By capturing front, left, and right views simultaneously, our dataset allows models to exploit complementary spatial and motion cues that are often ambiguous or occluded in frontal-only recordings, providing a stronger foundation for studying viewpoint-dependent sign understanding.
- Third, we provide empirical insights through extensive benchmarking. Our proposed multi-view fusion framework demonstrates that by leveraging three synchronized views, it achieves a 14.19-point absolute gain in BLEU-4 over the best single-view baseline.

2 Related Work

Sign Language Recognition and Translation Methods. Sign language understanding has attracted growing attention in recent years, ranging from isolated sign language recognition (ISLR), which classifies pre-segmented signs, to continuous sign language translation (CSLT), which requires modeling unsegmented sign streams and generating sentence-level translations. For CSLT, existing methods are generally divided into gloss-based and gloss-free approaches. Gloss-based methods use glosses as intermediate representations and typically decompose translation into sign-to-gloss

Table 1: **Comparison of representative continuous sign language datasets.** We compare datasets across language, scale, recording setting, multi-view availability, and domain/topic annotation.

Dataset	Language	Videos	Sentences	Vocab. sign	Frame rate	Signer ROI Size	Source	Multi. views	Domain/ Topic Anno.	Year
RWTH-Phoenix-Weather (Forster et al., 2012)	German	1,980	1,980	911	25	210×260	TV	✗	✗	2012
RWTH-Phoenix-2014T (Camgoz et al., 2018)	German	8,257	8,257	1,066	25	210×260	TV	✗	✗	2014
BOBSL (Albanie et al., 2021)	British	1,962	1,200,000	78,000	25	444×444	TV	✗	△	2021
OpenASL (Shi et al., 2022)	American	98,417	98,417	33,549	Varying	224×224	Web	✗	✗	2022
LSA-T (Dal Bianco et al., 2022)	Argentinian	14,880	14,254	14,239	30	224×224	Web	✗	✗	2022
Auslan-Daily (Shen et al., 2023)	Auslan	25,106	25,106	13,945	25/30	–	TV/Web	✗	△	2023
CE-CNSL (Zhu et al., 2024)	Chinese	5,988	5,988	3,515	Varying	256×256	Live	✗	✗	2024
SIGNUM (Von Agris & Kraiss, 2010)	German	33,210	780	450	30	776×578	Lab	✗	✗	2007
JSL (Bono et al., 2014)	Japanese	140	–	–	30	1280×720	Lab	✓	✗	2014
CSL (Huang et al., 2018)	Chinese	25,000	100	178	30	1280×720	Lab	✗	✗	2018
GSL (Adaloglou et al., 2021)	Greek	10,290	331	310	30	848×480	Lab	✗	✗	2019
KETI (Ko et al., 2019)	Korean	14,672	105	524	30	1920×1080	Lab	✗	✗	2019
How2Sign (Duarte et al., 2021)	American	35,191	35,191	16,000	30	1280×720	Lab	✓	△	2020
CSL-Daily (Zhou et al., 2021)	Chinese	20,654	8,193	2,000	30	1920×1080	Lab	✗	✗	2021
TriViS	**Vietnamese**	**84,453**	**12,000**	**14,133**	**30**	**1280×720**	**Lab**	✓	✓	**2026**

△ denotes datasets with coarse topic/category metadata, but without explicit sentence-level semantic domain annotation.

recognition and gloss-to-text generation (Camgoz et al., 2018; 2020; Chen et al., 2022; Kim et al., 2022). In contrast, gloss-free methods directly translate sign videos into text, avoiding the need for gloss annotations. Recent representative methods include GFSLT-VLP (Zhou et al., 2023), which performs visual-language pretraining with contrastive and masked language modeling objectives; GLoFE (Lin et al., 2023), which uses concept-level textual representations to guide cross-attention over visual features; and SpaMo (Hwang et al., 2025), which decomposes signs into spatial and motion components and integrates them with a large language model. While our dataset provides gloss annotations, our primary benchmark focuses on evaluating direct end-to-end sign-to-text translation without relying on intermediate gloss supervision, reflecting real-world scalable translation scenarios.

Datasets for Continuous Sign Language Recognition and Translation. Existing CSLR and CSLT datasets differ substantially in scale, annotation granularity, and recording conditions. Classical benchmarks such as SIGNUM (Von Agris & Kraiss, 2010), RWTH-Phoenix-Weather (Forster et al., 2012), and RWTH-PHOENIX-2014T (Camgoz et al., 2018) have played a central role in the development of continuous sign recognition and translation. More recent resources, including GSL (Adaloglou et al., 2021), CSL-Daily (Zhou et al., 2021), How2Sign (Duarte et al., 2021), OpenASL (Shi et al., 2022), Auslan-Daily (Shen et al., 2023), and BOBSL (Albanie et al., 2021), further improve vocabulary coverage, signer diversity, multimodal annotations, and real-world variability. Despite the progress enabled by recent CSLR and CSLT datasets, most existing benchmarks are primarily collected from frontal viewpoints, which may overlook important discriminative cues. In sign language, some signs can appear similar from a single angle but become clearly distinguishable when observed from a different viewpoint. To mitigate this issue, our dataset introduces multi-view sign recordings, providing complementary visual perspectives that support more robust recognition and translation of current methods.

3 The TriViS Dataset

3.1 Corpus Preparation and Multi-view Recording

As shown in Figure 1, the dataset construction begins with the selection of a suitable corpus. Our objective is to build a sufficiently large corpus whose gloss vocabulary comprehensively covers Vietnamese Sign Language while also being diverse in content and semantics. To this end, we drew from three complementary sources: a Vietnamese Sign Language Dictionary(Endangered Languages Project, 2025), Auslan Signbank(Cassidy et al., 2018) translated into Vietnamese, and Vietnamese learning materials. This source combination is designed to provide broad lexical coverage and reduce the bias toward a single linguistic domain. We then organized the corpus from a semantic perspective by assigning domain labels using the ChatGPT 5.4 API(OpenAI, 2026) with a few-shot prompting strategy. The generated labels were subsequently reviewed and manually validated to ensure annotation reliability. Overall, the corpus covers 10 semantic domains, with daily life being

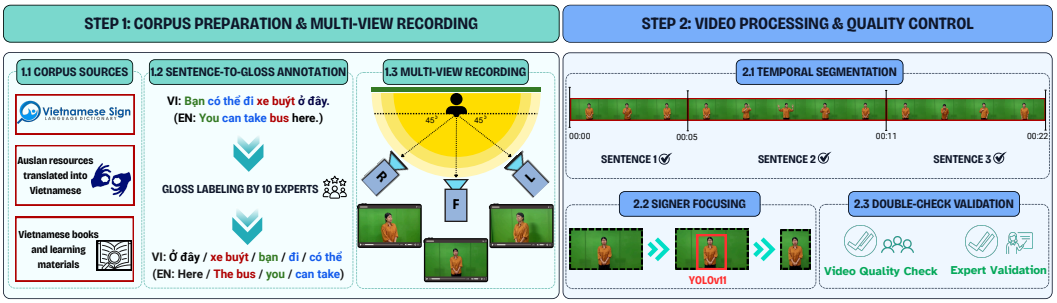


Figure 1: **Dataset construction pipeline.** TriViS is built through a two-stage workflow: corpus preparation with synchronized tri-view recording, followed by video processing and quality control.

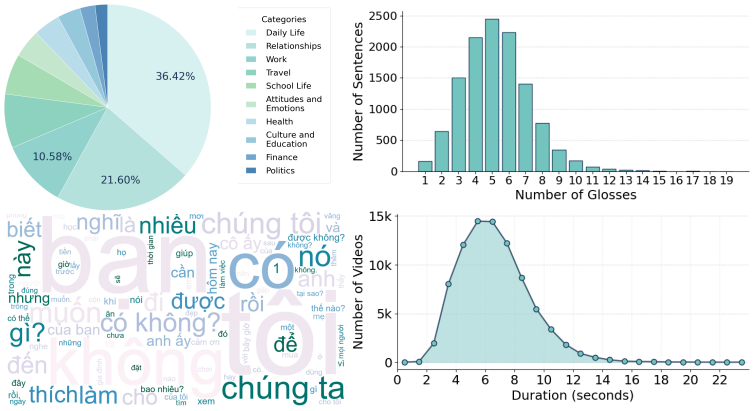
the most dominant category (36.42%), followed by relationships (21.60%) and work (10.58%). Other domains, including travel, school life, attitudes and emotions, and health, also account for notable portions of the dataset, as shown in Figure 2a.

Due to the fundamental differences between the grammatical structures of sign language and spoken language, the collected sentences could not be directly used in their original spoken-language form. To address this issue, we converted the spoken-language sentences into gloss sequences that better reflect sign language expression. This spoken-to-sign language conversion was conducted collaboratively by six deaf-education specialists from a reputable school for the deaf and four congenitally deaf signers, ensuring that the resulting annotations were both linguistically appropriate and consistent with actual signing practices. After the conversion was completed, the gloss sequences underwent cross-group verification, in which each group reviewed the annotations produced by the others to confirm their correctness. During this process, sentences that were overly complex or incompatible with grammar of Vietnamese Sign Language (VSL) were either removed or revised accordingly. This additional validation step helps improve annotation consistency and strengthens the agreement among different signers' interpretations, which is particularly important given that the same sentence can be expressed in different ways depending on the proficiency level of the signer.

A key feature of this dataset is the use of a three-camera recording setup. The recording involved four congenitally Deaf signers aged 29-51. They acquired VSL between the ages of 2 and 8 and had 22-43 years of VSL experience at the time of recording. All four were raised by hearing parents and are therefore categorized as Deaf-of-Hearing (detailed background in Appendix A.1). The signers participated in five recording sessions and wore different clothing across sessions to increase visual diversity. To capture signing performances from multiple viewpoints, the three-camera system consists of a left 45-degree view, a frontal view, and a right 45-degree view. Detailed signer metadata is provided in the Appendix. Throughout the recording process, the camera devices were securely mounted and carefully configured to ensure stability and consistency. OBS Studio (OBS Project) was then used to synchronously capture the three video streams corresponding to the three camera angles, ensuring temporal synchronization across all views. In addition, the timestamp of each sentence was documented using a built-in annotation tool during recording. This annotation supported accurate temporal segmentation and precise trimming of individual clips in the post-processing stage. As a result, the final dataset contains approximately 72k video clips recorded in a controlled laboratory environment with a green-screen background and additional lighting to enhance the brightness of the videos, along with around 12k video clips recorded in real-world settings with natural light conditions and background scenes. The diverse perspectives captured by this setup greatly enhance the utility of the dataset for various machine learning and computer vision research applications, enabling more accurate and robust recognition algorithms.

3.2 Video Processing and Quality Control

Starting from approximately 3,000 long-form recorded videos, we applied a multi-stage data processing and quality control pipeline to ensure data reliability. First, an automatic screening procedure was applied to detect technical issues, including abnormal video length, bit-level corruption, and file mismatches across the three views. Then, videos identified with frame rate inconsistencies were standardized to 30 fps to ensure temporal uniformity across the dataset, and each video was segmented into individual sentence clips according to the timestamps collected during the recording stage. To localize the signing region, YOLOv11 (Khanam & Hussain, 2024) was employed to extract



Characteristic	Value
# Sentences	12,000
# Video clips	84,453
# Views	3 (L, F, R)
Total duration (1 view)	51.92h
Total duration (all views)	155.77h
Gloss vocabulary size	14,133
Avg. sentence length	5.26 glosses
Sentence length range	1–19 glosses
Clip duration range	0.87–31.82 s
Avg. clip duration	6.64 s
# Signers	4
Sessions per signer	5
Video resolution	1280x720
Frame rate	30 FPS

Figure 2: **Overview of the TriViS dataset characteristics. (a)** Corpus statistics, including semantic domains, sentence lengths, word frequency, and clip duration. **(b)** Key dataset statistics, including scale, views, duration, vocabulary, signers, and video settings.

the human bounding box from each frame, and the videos were then spatially cropped based on the detected region in order to increase the visual focus on the signer.

The final stage of quality control consists of two phases. In the first phase, all processed videos were reviewed by a group of volunteers to identify issues such as timestamp misalignment, visual artifacts, and frame lag. In the second phase, 20% of the videos in the dataset were randomly sampled and sent to ten domain experts (six deaf-education specialists and four deaf signers) for verification of content accuracy and sign-language correctness, as shown in Figure 1.

3.3 Dataset Statistics and Analysis

In total, TriViS contains 12,000 sentences and 84,453 video clips captured from three synchronized views, with 51.92 hours of unique temporal content and 155.77 cumulative view-hours. The corpus contains 14,133 glosses and substantial variation in sentence and clip duration. Figure 2b summarizes the main dataset statistics, including signer, temporal, and recording characteristics.

4 Experiment Setup

Research Questions. The experiments evaluate the effectiveness of multi-view data for continuous sign language recognition and translation, together with ablations on cross-language difficulty, model generalization, view contribution, and LLM scale.

1. **Cross-language difficulty of continuous Vietnamese sign language.** We evaluate state-of-the-art vision-language APIs, including GPT-5 (Singh et al., 2025) and Gemini-2.5 (Comanici et al., 2025), on TriViS and other benchmarks in different languages. This experiment quantifies the performance gap of continuous Vietnamese sign language translation and highlights the need for dedicated Vietnamese resources.
2. **Multi-view vs. Single-view.** We compare conventional single-view inputs with multi-view settings using baselines such as SignCL (Ye et al., 2024), GFSLT-VLP (Zhou et al., 2023), GloFE (Lin et al., 2023), SpaMo (Hwang et al., 2025), LLaVA-Video-7B (Zhang et al., 2024), and Qwen3-VL (Bai et al., 2025). For the proposed SpaMo-based multi-view model, we evaluate both train 3-view/test 3-view and train 3-view/test 1-view settings to assess whether multi-view training improves representation learning and remains effective when only one view is available at inference.
3. **Effect of LLM scale and unseen-sentence generalization.** We compare LLMs with 1.5B, 3B, and 7B parameters under the same visual input setting using the proposed SpaMo-MV. Evaluation is conducted on both overlap and unseen-sentence protocols to measure in-distribution performance and generalization to novel sentence structures. We further compare single-view and three-view inputs to examine whether LLM scale and multi-view evidence provide complementary gains.

4. **Effectiveness of individual views.** We analyze the contribution of each camera viewpoint by training and evaluating models across the left, front, and right views under six cross-view configurations. This study identifies the most informative view and measures the degradation caused by train-test view mismatch.

Experiment Setting. We conduct experiments under two input settings: single-view and multiview-view. In the multiview-view setting, inspired by the computationally efficient design of SpaMo (Hwang et al., 2025), we develop a new three-view model, termed **SpaMo-MV**, which jointly takes synchronized videos from the left, front, and right cameras as input. The overall architecture is illustrated in Figure 3. For each view, spatial and motion features are first extracted offline using pretrained visual encoders. These features are then passed into a view-specific adapter, corresponding to the Left Adapter (LA), Front Adapter (FA), and Right Adapter (RA). The proposed model first learns view-specific representations and then fuses them through a lightweight attention pooling.

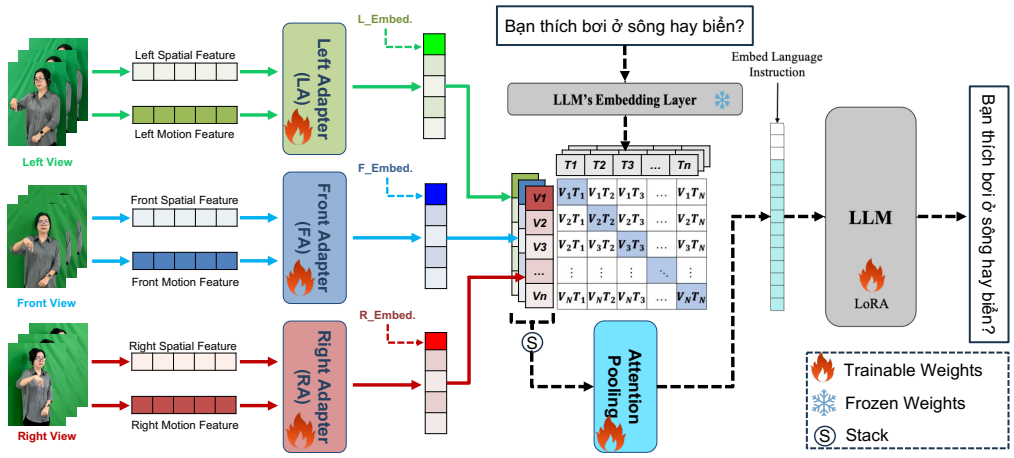


Figure 3: **Overall architecture of the proposed three-view framework SpaMo-MV.** The model extracts spatial and motion features from three synchronized camera views, maps them with view-specific adapters and learnable view embeddings, and fuses the resulting tokens through attention pooling before feeding them into an LoRA-adapted LLM decoder.

Let X_v denote the input video from view $v \in \{l, f, r\}$, corresponding to the left, front, and right cameras. For each view, we extract a spatial feature S_v and a motion feature M_v. The view-specific adapter $A_v(\cdot)$ maps these features into the LLM hidden space, and a learnable view embedding is added to encode the camera identity:

$$Z_v = A_v\left([S_v; M_v]\right) + v_Embed, \quad v \in \{l, f, r\}, \tag{1}$$

where $[S_v; M_v]$ denotes the combination of spatial and motion features, and v_Embed is the learnable embedding of view v. This design allows the model to preserve view-specific information while keeping all visual tokens in a shared representation space.

For three-view input, the adapted visual tokens are temporally aligned and stacked along the view dimension. A per-timestep attention pooling module is then applied to dynamically aggregate the three views:

$$U_t = \text{Stack}\left(Z_{l,t}, Z_{f,t}, Z_{r,t}\right), \quad V_t = \sum_{v \in \{l,f,r\}} \alpha_{t,v} Z_{v,t}, \quad \alpha_t = \text{Softmax}\left(U_t q\right), \tag{2}$$

where U_t contains the three view-specific tokens at time step t, q is a learnable attention query, and $\alpha_{t,v}$ denotes the attention weight assigned to view v. This formulation enables the model to emphasize the most informative camera view at each temporal position. The resulting sequence $V = \{V_1, V_2, \ldots, V_n\}$ serves as the fused multi-view visual representation.

The fused visual tokens are concatenated with the embedded language instruction and then fed into the LLM decoder Qwen-2.5.LoRA is applied to efficiently adapt the LLM to the sign language

recognition and translation task. The model is optimized with a combination of an autoregressive language modeling loss on the target text tokens and a CLIP-style alignment loss that encourages each view-specific visual representation to be close to its corresponding text embedding, thereby improving cross-modal consistency before multi-view fusion. For single-view training, the alignment loss is computed only for the available view.

All videos are processed using the same preprocessing pipeline before SpaMo/SpaMo-MV training, including FPS normalization, frame sampling, signer detection, and human-region cropping. Videos are resized to 224×224, with each input containing up to 512 frames. For SpaMo and SpaMo-MV, the per-GPU batch size is set to 12 with gradient accumulation of 2, resulting in an effective batch size of 24. These models are trained for up to 200 epochs using AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-8}$, and weight decay of 0.01. The base learning rate is 2×10^{-4}, with differential learning-rate multipliers for the LLM-LoRA branch, side-view adapters, and view-fusion.

All experiments use a single NVIDIA A100-SXM4 80 GB GPU. SignCL and GFSLT-VLP are trained in full precision (`fp32`), while SpaMo and SpaMo-MV are trained in mixed precision (`bf16`). SignCL uses a batch size of 16 and is trained for 80 epochs with AdamW and cosine learning-rate scheduling, taking approximately 100 minutes per epoch, corresponding to about 133 wall-clock hours for a full run. GFSLT-VLP follows a two-stage pipeline with 80 epochs of VLP pre-training and 80 epochs of SLT fine-tuning using cached visual features; raw-video multiview attempts are excluded because they fail to complete one epoch due to excessive decoding time and memory cost. For SpaMo, the single-view VideoLLaMA3-7B model takes approximately 20 minutes per epoch, requiring about 26.6 hours for 80 epochs, while SpaMo-MV with three synchronized views takes approximately 52 minutes per epoch, requiring about 69 hours for 80 epochs. The lighter SpaMo-MV variant with VideoLLaMA3-2B requires approximately 6.4 minutes per epoch and about 8.5 hours for 80 epochs. Overall, training time varies mainly with model scale, number of input views, and whether cached features or raw-video decoding are used.

Evaluation Metrics. We adopt four standard metrics: BLEU-1 and BLEU-4 (Post, 2018) for n-gram overlap, ROUGE-L (Lin, 2004) for longest common subsequence matching, and WER (Morris et al., 2004) for word-level error rate. Higher BLEU and ROUGE-L indicate better performance, while lower WER is preferred.

5 Evaluation

5.1 Cross-language Difficulty of Continuous Vietnamese Sign Language

To assess whether continuous Vietnamese sign language translation poses a greater challenge than other languages under current models, we evaluate two state-of-the-art vision-language APIs, GPT-5 (Singh et al., 2025) and Gemini-2.5 (Comanici et al., 2025), on TriViS and four established benchmarks: Auslan-Daily (Shen et al., 2023), How2Sign (Duarte et al., 2021), CSL-CE (Zhou et al., 2021), and PHOENIX-2014T (Camgoz et al., 2018). For each dataset, 1,000 sentences are randomly sampled and evaluated under few-shot inference. As shown in Table 2, Vietnamese consistently exhibits the weakest performance across both APIs. Gemini-2.5 obtains a BLEU-1 of only 5.59 on TriViS, compared to 20.18 on Auslan-Daily and 11.60 on How2Sign, while the WER on TriViS reaches 78.40, far exceeding all other datasets. This performance gap underscores two primary insights: current vision-language models remain insufficiently grounded for continuous sign language translation, and Vietnamese Sign Language (VSL) appears particularly disadvantaged. Although this disparity aligns with the acute shortage of continuous VSL pretraining resources, it may also reflect dataset difficulty or domain differences.

Table 2: **Few-shot API evaluation across sign language datasets.** Underline denotes the worst performance.

Dataset	API	B-1↑	B-4↑	R-L↑	WER↓
Auslan-Daily (Shen et al., 2023) (Australian)	GPT-5	16.30	4.35	30.80	58.20
	Gemini-2.5	20.18	5.70	31.25	55.90
How2Sign (Duarte et al., 2021) (American)	GPT-5	12.80	3.10	29.40	60.80
	Gemini-2.5	11.60	2.70	27.85	63.10
CSL-CE (Zhou et al., 2021) (Chinese)	GPT-5	10.30	2.35	27.10	65.20
	Gemini-2.5	9.70	2.10	26.40	66.50
PHOENIX-2014T (Camgoz et al., 2018) (German)	GPT-5	10.67	2.65	28.80	62.70
	Gemini-2.5	9.25	2.30	27.50	64.00
TriViS (Vietnamese)	GPT-5	11.92	2.44	28.24	<u>78.40</u>
	Gemini-2.5	<u>5.59</u>	<u>1.40</u>	<u>21.65</u>	72.60

Regardless of the underlying factors, these results highlight continuous VSL translation as an under-addressed task, strongly motivating dedicated benchmarks such as TriViS.

5.2 MULTI-VIEW VS. SINGLE-VIEW

Table 3: **Comparison between single-view and multi-view settings.** All methods are evaluated on laboratory and real-world subsets. The SpaMo-MV model is tested under two configurations: train multi-view/test front-view and train multi-view/test multi-view. **Bold** denotes the best performance.

Method	Lab				Real-world			
	BLEU-1 ↑	BLEU-4 ↑	ROUGE-L ↑	WER ↓	BLEU-1 ↑	BLEU-4 ↑	ROUGE-L ↑	WER ↓
Front-view training								
SignCL (Ye et al., 2024)	10.45	0.77	3.59	116.20	9.78	0.64	3.18	117.40
GFSLT-VLP (Zhou et al., 2023)	9.02	0.63	3.91	118.20	9.36	0.48	3.93	121.58
GloFE (Lin et al., 2023)	20.26	5.62	32.57	92.15	20.20	**6.03**	**29.41**	95.98
Qwen3-VL (Bai et al., 2025)	9.86	1.91	28.24	112.17	8.59	2.03	26.94	111.84
LLaVA-Video-7B (Zhang et al., 2024)	15.26	3.62	25.92	98.26	11.42	1.54	5.65	**92.76**
SpaMo (Hwang et al., 2025)	40.12	11.09	30.25	94.93	13.42	1.00	18.05	128.40
Multi-view training								
SpaMo-MV (Hwang et al., 2025) (test front-view)	40.43	12.26	**33.79**	96.71	12.92	1.54	14.63	108.55
SpaMo-MV (Hwang et al., 2025) (test multi-view)	**44.82**	**25.28**	32.49	**87.54**	**22.58**	2.60	19.84	102.34
Improv. over best single-view	+4.70	+14.19	-0.08	-4.61	+2.38	-3.43	-9.57	+9.58

Table 3 presents the comparison between single-view and multi-view settings on both the laboratory (Lab) and real-world subsets. Overall, the SpaMo-MV model trained and tested with three views achieves the best performance across most metrics. Specifically, on the Lab subset, SpaMo-MV (test multi-view) obtains a BLEU-4 score of 25.28, yielding a substantial improvement of +14.19 points over the best single-view method, while also reducing WER to 87.54 (-4.61 compared to single-view). These results demonstrate that integrating visual information from multiple viewpoints enables the model to capture complementary spatial cues that a single viewpoint alone cannot provide, which is consistent with the findings of Dinh et al. (Dinh et al., 2025) on word-level sign language.

Another noteworthy observation concerns the stability of different methods when transitioning from the Lab to the Real-world environment. GloFE (Lin et al., 2023), a skeleton-based method, maintains relatively stable performance across the two settings (BLEU-4: 5.62 $\rightarrow$ 6.03; ROUGE-L: 32.57 $\rightarrow$ 29.41). In contrast, RGB-based methods such as SpaMo (Hwang et al., 2025) exhibit a pronounced degradation on real-world data (BLEU-1: 40.12 $\rightarrow$ 13.42). This suggests that skeleton-based features are more invariant to changes in background, lighting, and other environmental factors, whereas RGB-based features are more susceptible to distribution shift between controlled and real-world conditions. Even when the model is trained with three views, as in SpaMo-MV, this degradation remains substantial (BLEU-1: 44.82 $\rightarrow$ 22.58), although it still outperforms single-view counterparts. This constitutes a significant limitation of current RGB-based models for sign language under real-world deployment scenarios.

5.3 EFFECT OF LLM SCALE AND SENTENCE GENERALIZATION

To investigate how language model capacity interacts with multi-view visual input, we evaluate three LLM backbone sizes (1.5B, 3B, 7B) under both overlap and unseen-sentence protocols. As illustrated in Figure 4, scaling the LLM yields modest gains under the single-view setting (BLEU-4: 9.04 $\rightarrow$ 11.09), but produces substantially larger improvements when combined with three-view input (BLEU-4: 22.71 $\rightarrow$ 25.28). This amplified gain can be attributed to the larger hidden dimension of the decoder: Qwen2.5-1.5B operates with a hidden size of 1,536, whereas Qwen2.5-7B uses 3,584, providing a richer representational space for projecting fused visual tokens.

The figure also reveals a striking generalization gap. With single-view 7B, BLEU-4 drops from 11.09 (overlap) to merely 1.64 (unseen). Three-view input dramatically narrows this gap: under the same 7B backbone, BLEU-4 on unseen sentences reaches 11.96, nearly a sevenfold improvement over the single-view counterpart. This confirms that multi-view observations provide richer visual evidence that helps the model generalize beyond memorized sentence patterns.

Table 4 shows that single-view training performs best when the training and test views are matched, as indicated by the stronger diagonal results. For example, training on the front view achieves the highest single-view performance on the front test view, while left-view and right-view training similarly perform best on their corresponding test views. In contrast, cross-view evaluation leads to a performance drop, suggesting that models trained from one angle do not generalize well to unseen viewpoints. This highlights the view-dependent nature of sign-language recognition in our setting.

Training with multiple views provides more stable performance across test views. Although the three-view model does not always dominate every metric, it achieves consistently strong results on each view, with particularly large gains in BLEU-1 and ROUGE-L compared with most cross-view single-view settings. These results suggest that multi-view training exposes the model to richer visual variation, reducing sensitivity to camera viewpoint and improving robustness across different views.

5.4 Effectiveness of Individual Views

Table 4: **Cross-view evaluation of individual camera views.** Models are trained on one view and evaluated on different camera views. For each view, we report BLEU-1, BLEU-4, and ROUGE-L.

Train View	Front View			Left View			Right View		
	BLEU-1	BLEU-4	ROUGE-L	BLEU-1	BLEU-4	ROUGE-L	BLEU-1	BLEU-4	ROUGE-L
Front	40.12	11.09	30.25	20.41	3.16	24.00	18.55	2.38	23.15
Left	20.23	2.99	25.74	25.60	**10.68**	**29.09**	20.68	2.31	25.65
Right	21.30	3.66	24.22	20.29	2.85	23.49	25.03	**10.56**	27.85
Three-view	**40.43**	**12.26**	**33.79**	**37.61**	8.28	28.37	**38.88**	9.79	**31.41**

6 Conclusion

In this paper, we introduced TriViS, a large-scale tri-view dataset for continuous Vietnamese sign language recognition and translation. The dataset contains 12,000 sentences and 84,453 video clips recorded from three synchronized camera views: left, front, and right. To support reliable model training and evaluation, TriViS was constructed through a complete pipeline that integrates corpus creation, gloss conversion, multi-view recording, video preprocessing and quality validation.

To assess the usefulness of the proposed dataset, we conducted extensive experiments to evaluate the role of multi-view information in CSLT. Results show that multi-view inputs provide complementary visual cues and consistently improve performance over single-view settings. In addition, the study on sentence-independent splits further reveals the challenge of generalizing to unseen sentence content. Cross-view evaluation also shows that different camera views contribute differently to recognition performance, while the LLM-size analysis highlights the trade-off between generation quality and computational cost.

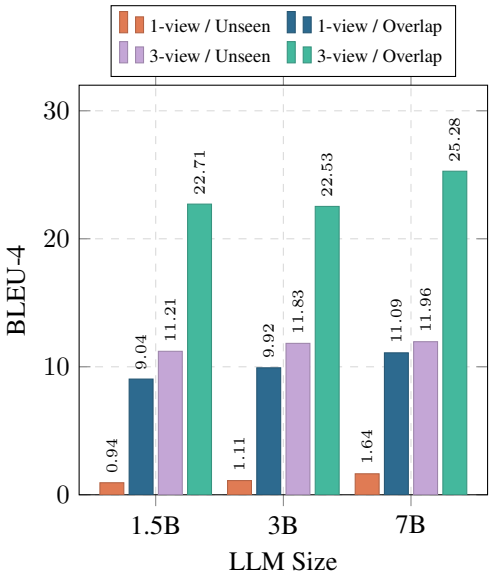


Figure 4: **BLEU-4 across LLM scales and view configurations.**

Overall, TriViS provides a new benchmark for studying multi-view continuous sign language understanding in Vietnamese. These findings indicate that multi-view data is important for improving visual representation and robustness, particularly when sign motions are ambiguous from a single viewpoint. Building on this benchmark, future work will focus on more effective multi-view fusion strategies, stronger adaptation of visual features to language models, and more robust recognition under practical settings where only partial views are available.

7 Limitation and Future Work

We identify several key limitations of the current dataset and framework. First, the current release includes only four signers, which limits signer-independent generalization and does not capture the full individual and regional variation of VSL. Generalization may also be affected by signer background, age of VSL acquisition, and recording conditions. In addition, the synchronized three-camera setup is less practical for real-time consumer deployment. However, multi-view training can still benefit single-camera inference, and the tri-view design also enables broader research directions such as cross-view rendering and multi-view 3D reconstruction. Future work will expand signer and environment diversity and investigate robust single- and missing-view inference.

AI USE STATEMENT

We used generative AI tools solely to support manuscript polishing, including improvements to organization, presentation, and the clarity of experimental reporting. These tools were not used to define the research problem, develop the core technical contributions, or design the experiments. All AI-assisted content was carefully reviewed and verified by the authors for accuracy, consistency, and fidelity to the intended technical meaning. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Nikolas Adaloglou, Theocharis Chatzis, Ilias Papastratis, Andreas Stergioulas, Georgios Th Papadopoulos, Vassia Zacharopoulou, George J Xydopoulos, Klimnis Atzakas, Dimitris Papazachariou, and Petros Daras. A comprehensive study on deep learning-based methods for sign language recognition. *IEEE transactions on multimedia*, 24:1750–1762, 2021.

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. Bbc-oxford british sign language dataset. Technical report, University of Oxford, 2021.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Mayumi Bono, Kouhei Kikuchi, Paul Cibulka, and Yutaka Osugi. A colloquial corpus of japanese sign language: Linguistic resources for observing sign language conversations. In *LREC*, pp. 1898–1904, 2014.

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

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Steve Cassidy, Onno Crasborn, Henri Nieminen, Wessel Stoop, Micha Hulsbosch, Susan Even, Erwin Komen, and Trevor Johnston. Signbank: Software to support web based dictionaries of sign language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 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)*, pp. 5120–5130, 2022.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Pedro Dal Bianco, Gastón Ríos, Franco Ronchetti, Facundo Quiroga, Oscar Stanchi, Waldo Hasperué, and Alejandro Rosete. Lsa-t: The first continuous argentinian sign language dataset for sign language translation. In *Ibero-American Conference on Artificial Intelligence*, pp. 293–304. Springer, 2022.

Nguyen Son Dinh, Tuan Dung Nguyen, Duc Tri Tran, Nguyen Dang Huy Pham, Thuan Hieu Tran, Ngoc Anh Tong, Quang Huy Hoang, and Phi Le Nguyen. Sign language recognition: A large-scale multi-view dataset and comprehensive evaluation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7887–7897. IEEE, 2025.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2sign: a large-scale multimodal dataset for continuous american sign language. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2734–2743. IEEE, 2021.

Endangered Languages Project. Vietnamese Sign Language Dictionary. https://endangeredlanguages.com/resource/vietnamese-sign-language-dictionary, 2025. Accessed: 2026-04-29.

Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus H Piater, and Hermann Ney. Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus. In *LREC*, volume 9, pp. 3785–3789, 2012.

Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. Video-based sign language recognition without temporal segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Eui Jun Hwang, Sukmin Cho, Junmyeong Lee, and Jong C Park. An efficient sign language translation using spatial configuration and motion dynamics with llms. In *NAACL*, 2025.

Nada B Ibrahim, Hala H Zayed, and Mazen M Selim. Advances, challenges and opportunities in continuous sign language recognition. *Journal of Engineering and Applied Sciences*, 15(5): 1205–1227, 2020.

Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.

Minchae Kim, Jungeun Kim, and Ha Young Kim. Sign2gloss2text-based sign language translation with enhanced spatial-temporal information centered on sign language movement keypoints. *Journal of Korea Multimedia Society*, 25(10):1535–1545, 2022.

Edward S Klima and Ursula Bellugi. *The signs of language*. Harvard University Press, 1979.

Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. Neural sign language translation based on human keypoint estimation. *Applied sciences*, 9(13):2683, 2019.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. 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)*, pp. 12904–12916, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.722.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Proc. Interspeech 2004*, pp. 2765–2768, 2004.

OBS Project. OBS Studio. URL https://obsproject.com/. Free and open-source software for video recording and live streaming.

OpenAI. GPT-5.4 Model. https://developers.openai.com/api/docs/models/gpt-5.4, 2026. Accessed: 2026-04-29.

Matt Post. A call for clarity in reporting bleu scores. In *Proceedings of the third conference on machine translation: Research papers*, pp. 186–191, 2018.

Wendy Sandler, Diane Carolyn Lillo-Martin, Diane Lillo-Martin, and Wendy Sandler. *Sign language and linguistic universals*, volume 1. Cambridge university press Cambridge, 2006.

Xin Shen, Shaozu Yuan, Hongwei Sheng, Heming Du, and Xin Yu. Auslan-daily: Australian sign language translation for daily communication and news. *Advances in Neural Information Processing Systems*, 36:80455–80469, 2023.

Bowen Shi, Diane Brentari, Gregory Shakhnarovich, and Karen Livescu. Open-domain sign language translation learned from online video. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6365–6379, 2022.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

William C Stokoe Jr. Sign language structure: An outline of the visual communication systems of the american deaf. *Journal of deaf studies and deaf education*, 10(1):3–37, 2005.

Ulrich Von Agris and Karl-Friedrich Kraiss. Signum database: Video corpus for signer-independent continuous sign language recognition. In *4th workshop on the representation and processing of sign languages: corpora and sign language technologies*, pp. 243–246, 2010.

Jinhui Ye, Xing Wang, Wenxiang Jiao, Junwei Liang, and Hui Xiong. Improving gloss-free sign language translation by reducing representation density. *Advances in Neural Information Processing Systems*, 37:107379–107402, 2024.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. Gloss-free sign language translation: Improving from visual-language pretraining. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20814–20824. IEEE, 2023.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 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)*, 2021.

Qidan Zhu, Jing Li, Fei Yuan, Jiaojiao Fan, and Quan Gan. A chinese continuous sign language dataset based on complex environments. *arXiv preprint arXiv:2409.11960*, 2024.

A APPENDIX

A.1 SIGNER METADATA

To provide additional context on signer diversity, we report anonymized demographic and linguistic background information for all four signers in Table 5. Regional information is reported only at an aggregate level to protect participant privacy. The signers also have educational backgrounds related to Special Education, Deaf education, VSL interpreting, or Art Education, providing substantial familiarity with VSL use and Deaf communication contexts.

Table 5: Demographic and linguistic background of the TriViS signers. DoH denotes Deaf-of-Hearing, referring to Deaf individuals raised by hearing parents.

Signer	**Gender**	**Age**	**VSL Acq. Age**	**VSL Exp.**	**Educational Background**	**Deaf Background**
Signer 1	Female	32	2	30 years	University degree; Training in VSL interpreting and Deaf education	Congenitally Deaf; DoH
Signer 2	Female	33	6	27 years	College degree in Special Education; University degree in Art Education	Congenitally Deaf; DoH
Signer 3	Female	29	7	22 years	College degree in Special Education	Congenitally Deaf; DoH
Signer 4	Male	51	8	43 years	College degree in Special Education	Congenitally Deaf; DoH

A.2 SIGNER-INDEPENDENT EVALUATION

The original data split contains the same signer identities across training and test sets and is primarily designed to study viewpoint effects under controlled conditions. To further evaluate generalization to unseen signers, we conduct a four-fold leave-one-signer-out evaluation using SpaMo. In each fold, all samples from one signer are held out for testing, while data from the remaining three signers are used for training. The model configuration and training protocol are kept unchanged across all folds.

Table 6: Leave-one-signer-out evaluation using SpaMo. Each fold evaluates the model on a signer unseen during training. "Original" denotes the original random split with mixed signers in the training and test sets.

Setting	Test Signer	BLEU-1 ↑	BLEU-4 ↑	ROUGE-L ↑	WER ↓
Fold 1	Signer 1	19.26	6.04	18.62	120.11
Fold 2	Signer 2	15.92	5.50	26.62	145.86
Fold 3	Signer 3	20.27	4.28	28.91	119.83
Fold 4	Signer 4	15.11	4.19	22.60	150.91
Mean $\pm$ std	Unseen signer	17.64 $\pm$ 2.51	5.00 $\pm$ 0.91	24.19 $\pm$ 4.54	134.18 $\pm$ 16.53
Original	Mixed signers	40.12	11.09	30.25	94.93

The leave-one-signer-out results show a substantial performance gap relative to the original mixed-signer split, confirming that generalization to unseen signers remains challenging. BLEU-4 is comparatively consistent across the four held-out signers, whereas WER exhibits larger variation, suggesting that signer-specific characteristics continue to affect recognition and translation performance. These results further motivate expanding TriViS with additional signers and more diverse signer backgrounds in future releases.