An Efficient Sign Language Translation Using Spatial Configuration and Motion Dynamics with LLMs

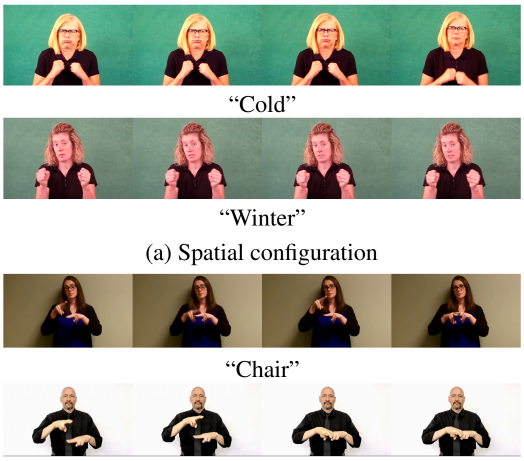
Anonymous ACL submission

Abstract

Gloss-free Sign Language Translation (SLT) converts sign videos directly into spoken language sentences without relying on glosses. Recently, Large Language Models (LLMs) have shown remarkable translation performance in gloss-free methods by harnessing their powerful natural language generation capabilities. However, these methods often rely on domain-specific fine-tuning of visual encoders to achieve optimal results. By contrast, this paper emphasizes the importance of capturing the spatial configurations and motion dynamics inherent in sign language. With this in mind, we introduce **Spa**tial and **Mo**tion-based Sign Language Translation (**SpaMo**), a novel LLM-based SLT framework. The core idea of SpaMo is simple yet effective. We first extract spatial and motion features using off-the-shelf visual encoders and then input these features into an LLM with a language prompt. Additionally, we employ a visual-text alignment process as a warm-up before the SLT supervision. Our experiments demonstrate that SpaMo achieves state-of-the-art performance on two popular datasets, PHOENIX14T and How2Sign[1].

1 Introduction

Sign language is a visual means of communication primarily used by Deaf communities, relying on physical movements rather than spoken words. In this paper, we tackle Sign Language Translation (SLT), focusing on converting sign videos into spoken language sentences. Early SLT methods (Camgoz et al., 2020; Voskou et al., 2021; Zhou et al., 2021b,a; Yin et al., 2021; Jin et al., 2022; Chen et al., 2022a,b; Zhang et al., 2023b) have primarily relied on *glosses*—written representations of signs using corresponding words. Glosses provide a structured form of sign language, which helps identify semantic boundaries within continuous sign sequences. This, in turn, allows the models to better comprehend the overall content of the sign videos (Yin et al., 2023; Wei and Chen, 2023). However, annotating glosses is a labor-intensive and time-consuming process that requires expertise in sign language. This significantly hinders the expansion of sign language datasets and limits the development of SLT methods (Li et al., 2020b; Shi et al., 2022; Lin et al., 2023).


“Cold”

“Winter”

(a) Spatial configuration

“Chair”

“Sit”

(b) Motion dynamics

Figure 1: Visual examples of spatial configurations and motion dynamics in sign language. The images are sourced from WLASL (Li et al., 2020a).

To address these limitations, there is a shift towards gloss-free methods that rely solely on the sign videos and corresponding translated text. While these methods still underperform compared to the gloss-based methods, efforts have been made to reduce the performance gap by focusing on temporal semantic structures (Li et al., 2020b) and aligning visual and textual modalities (Zhao et al., 2021; Yin et al., 2023; Fu et al., 2023; Zhou et al., 2023). Recently, Large Language Models (LLMs) have demonstrated remarkable translation performance in a gloss-free setting by harnessing their powerful language generation capabilities. How-

[1] Code will be available at https://anonymous.4open.science/r/SpaMo-9CB4/

ever, the modality gap between the continuous sign videos and discrete text poses a challenge for the LLMs in effectively understanding the sign videos. To address this problem, these methods fine-tune their visual encoders to be more domain-specific (Wong et al., 2024; Chen et al., 2024; Rust et al., 2024; Gong et al., 2024). However, these approaches are resource-intensive and time-consuming, presenting significant challenges that are hard to overcome. Moreover, there has been limited research on how the LLMs process the sign videos and the reasons behind their superior performance compared to other methods.

In this paper, we challenge the understanding that fine-tuning visual encoders is necessary to achieve optimal performance in gloss-free SLT based on LLMs. Instead, we revisit a fundamental question: *"What are the key components in sign language that most effectively convey meaning?"* Our focus is on the important roles of **spatial configuration** and **motion dynamics** in sign language. Spatial configuration encompasses the arrangement and positioning of signs within the signing space, including hand shapes, facial expressions, and body postures. These components work together to distinguish different signs and convey their intended meanings (Emmorey and Casey, 1995). As shown in Figure 1a, the signs for "cold" and "winter" both use the same handshape, with a shivering motion of the fists. The primary difference lies in the facial expressions: "cold" is typically accompanied by a tensed or grimaced expression, while "winter" may feature a more neutral expression. Motion dynamics, on the other hand, involve the path, speed, and rhythm of hand movements, illustrating how movements alter the meanings of signs over time (Bosworth et al., 2019). As shown in Figure 1b, the signs for "chair" and "sit" both use the same 'H" handshape and involve the interaction of both hands. However, the motion differentiates these signs: "chair" involves a repetitive tapping motion, while "sit" involves a single, smooth motion. These examples highlight the importance of both spatial configuration and motion dynamics in conveying accurate messages in sign language.

To this end, we introduce a novel gloss-free framework, **Spa**tial and **Mo**tion-based Sign Language Translation (**SpaMo**), designed to fully leverage the spatial configurations and motion dynamics in the sign videos, all without the need for domain-specific fine-tuning. As shown in Figure 2, the core idea is simple: We extract spatial and motion features using two different visual encoders and input these features into an LLM with a language prompt. Specifically, we use a pre-trained image encoder (e.g., ViT) as **Spatial Encoder (SE)**, to individually encode each frame for its spatial features. To further refine the spatial configurations, we apply S^2 scaling (Shi et al., 2024), which processes a sign image at multiple scales. Additionally, we use a video encoder (e.g., VideoMAE) as **Motion Encoder (ME)** to encode sign clips (groups of sign frames) into the motion features. To capture finer motion dynamics, we apply a sliding window approach, which results in implicit gloss-level representations (Cheng et al., 2023; Hwang et al., 2024). Next, **Sign Adapter (SA)**, comprising Multi-Layer Perceptron (MLP) layers, transfers these features to the LLM. Additionally, we introduce **Visual-Text Alignment (VT-Align)**, a training strategy to effectively narrow the modality gap, ensuring efficient training and enhanced translation performance.

In all, our contributions can be summarized as:

- We introduce SPaMo, a novel gloss-free framework. Our method eliminates the need for fine-tuning visual encoders by utilizing readily available, off-the-shelf models. Rather than focusing on the expensive training, it focuses on the fundamental components in sign language, enabling offering a simple yet effective translation with LLMs.

- Our proposed method achieves state-of-the-art performance on two popular sign language datasets: PHOENIX14T and How2Sign.

- We provide a novel and comprehensive analysis of how the LLM interprets the sign videos within its embedding space and translate them into corresponding text.

2 Related Work

2.1 Gloss-free Sign Language Translation

Gloss-free SLT directly converts sign videos into spoken language sentences without relying on glosses. These methods, however, often underperform compared to gloss-based methods (Camgoz et al., 2020; Voskou et al., 2021; Zhou et al., 2021b,a; Yin et al., 2021; Jin et al., 2022; Chen et al., 2022a,b; Zhang et al., 2023b). To address the performance gap, recent work has focused on several key areas: enhancing the temporal semantic structure (Li et al., 2020b), improving the alignment between visual and textual modalities (Zhao et al., 2021; Lin et al., 2023; Fu et al., 2023), leveraging large language models (LLMs) (Wong et al., 2024; Gong et al., 2024; Chen et al., 2024), and scaling efforts by utilizing larger sign language datasets (Uthus et al., 2024; Rust et al., 2024).

Despite these advancements, most gloss-free methods depend on fine-tuning visual encoders using the glosses (Li et al., 2020b; Yin et al., 2023; Fu et al., 2023), target translations (Zhou et al., 2023; Wong et al., 2024), or self-supervised learning (Gong et al., 2024; Rust et al., 2024). In particular, fine-tuning with the glosses (gloss-supervision) helps the visual encoders to be more domain-specific training on continuous or isolated Sign Language Recognition (SLR) datasets, such as WLASL (Li et al., 2020a) and PHOENIX14T (Camgoz et al., 2018). Consequently, we classify these methods as *weakly gloss-free* due to the implicit involvement of gloss information, as further elaborated in Section 4.3. On the other hand, the rest of the fine-tuning methods eliminate reliance on these annotations. However, these methods often require substantial resources, making it difficult to achieve robust visual representations and enhance translation performance without access to a sufficiently large dataset. To address this limitation, our approach diverges from this norm by focusing on capturing the spatial configurations and motion dynamics, thereby avoiding the need for resource-intensive fine-tuning.

2.2 Large Language Models

Recently, LLMs (Touvron et al., 2023; Chiang et al., 2023; Chung et al., 2024) have demonstrated impressive text generation capabilities, through extensive training on web-scale text corpora. This extensive training has endowed the LLMs with robust generalization abilities across various tasks. Notable applications include multilingual translation (Zhu et al., 2023; Zhang et al., 2023a; Gao et al., 2024), pose generation (Feng et al., 2024; Zhang et al., 2024a), and, visual question answering (Li et al., 2023; Liu et al., 2024a,b). In SLT, LLMs also demonstrate impressive translation performance. These methods focus on aligning high-dimensional visual features with inputs comprehensible to LLMs. To achieve this alignment, the visual encoders are pre-trained to produce language-like tokens (Gong et al., 2024), utilize pseudo-glosses (Wong et al., 2024), or perform video-grounded text generation tasks (Chen et al., 2024).

Our work shows that a domain-specific fine-tuning of the visual encoders is unnecessary for optimal performance. Instead, we extract spatial and motion features and pass them to the LLM through a simple connector, accompanied by a light warm-up process. This approach is both simple and effective, proving that a complex learning process is not required to achieve peak performance.

3 Method

We first give an overview of our framework in Section 3.1, We then explain SE and ME in Sections 3.2 and 3.3, respectively. Next, we discuss SA in Section 3.4 and VT-Align in Section 3.5. Finally, we explain the training details in Section 3.6.

3.1 Framework Overview

Given a sign video $X = \{x_i\}_{i=1}^{T}$, where each frame $x_i \in \mathbb{R}^{H\times W}$ represents a frame with height H and width W, the objective of SLT is to generate a corresponding spoken language sentence $Y = \{y_j\}_{j=1}^{U}$, composed of U words. Previous gloss-free methods (Zhou et al., 2023; Wong et al., 2024; Gong et al., 2024; Chen et al., 2024) have involved fine-tuning visual encoders using sign language data to be more domain-specific, thereby improving on translation performance. However, while this fine-tuning process injects more domain knowledge at the feature extraction level, it is often unnecessary and resource-intensive, especially with LLMs, which already maintain rich visual information from the visual encoder in their latent space (Zhang et al., 2024b). This creates a trade-off, but we argue that the latter approach is more effective. Thus, we emphasize the importance of encoding the spatial configurations and motion dynamics in sign language and decoding this information through proper alignment and training.

As shown in Figure 2, Spatial Encoder (SE) and Motion Encoders (ME) extract two distinct features from the sign video X: Spatial features Z_s capture the spatial configurations (Emmorey and Casey, 1995), and motion features Z_m represent the motion dynamics (Bosworth et al., 2019). These features are then integrated into a combined sign feature Z_{sm} via Sign Adapter (SA). The combined feature is then fed to an LLM with an language prompt, guiding the LLM to generate the translation in the desired language. Additionally, we perform Visual-Text Alignment (VT-Align) to minimize the gap between the visual and textual modalities before and during training under SLT supervision.

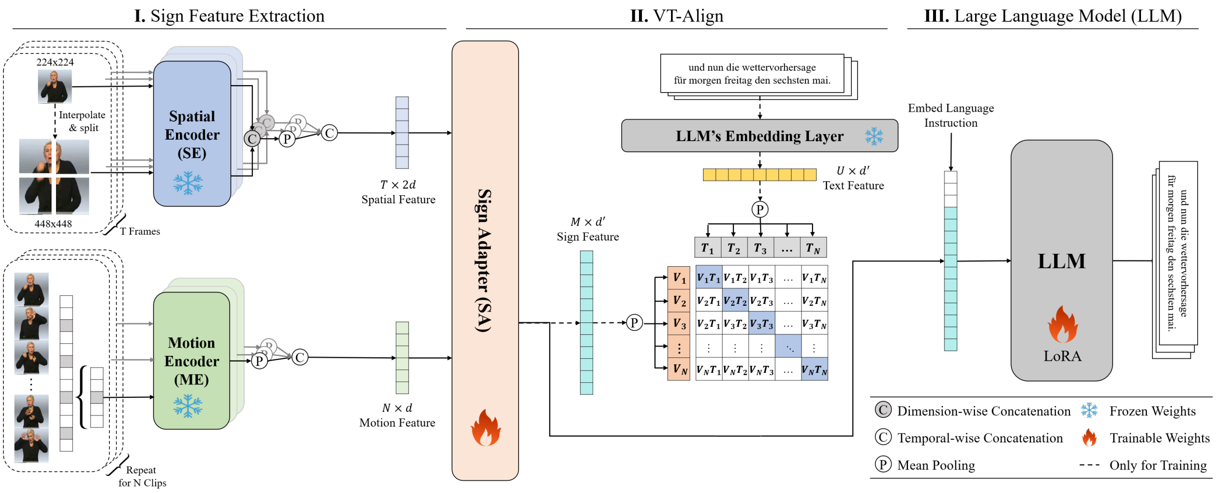

I. Sign Feature Extraction
II. VT-Align
III. Large Language Model (LLM)
224x224
Interpolate & split
448x448
T Frames
Spatial Encoder (SE)
C
P
$T \times 2d$
Spatial Feature
Motion Encoder (ME)
$N \times d$
Motion Feature
Repeat for N Clips
Sign Adapter (SA)
$M \times d'$
Sign Feature
und nun die wettervorhersage für morgen freitag den sechsten mai.
LLM's Embedding Layer
$U \times d'$
Text Feature
T_1 T_2 T_3 … T_N
V_1 V_2 V_3 ⋮ V_N
V_1T_1 V_1T_2 V_1T_3 … V_1T_N
V_2T_1 V_2T_2 V_2T_3 … V_2T_N
V_3T_1 V_3T_2 V_3T_3 … V_3T_N
V_NT_1 V_NT_2 V_NT_3 … V_NT_N
Embed Language Instruction
LLM
LoRA
und nun die wettervorhersage für morgen freitag den sechsten mai.
C Dimension-wise Concatenation
C Temporal-wise Concatenation
P Mean Pooling
Frozen Weights
Trainable Weights
Only for Training

Figure 2: An overview of the SpaMo framework, which consists of three parts: (i) **Sign Feature Extraction**: Spatial and motion features are extracted using SE and ME, leveraging the S^2 and sliding window approaches to capture detailed spatial configurations and motion dynamics. (ii) **VT-Align**: The extracted features are combined within SA to form a unified sign feature. During training, a warm-up process is employed to ensure that SA has well-initialized weights, effectively bridging the modality gap between the sign video and text. (iii) **LLM**: Finally, the LLM processes the sign feature along with a language-instructive prompt and is trained using LoRA.

3.2 Spatial Encoder

SE extracts spatial features Z_s from the sign video X. We utilize a pre-trained image encoder (e.g., ViT), which is kept frozen, and enhances its capability to capture more detailed spatial information by applying Scaling on Scales (S^2) (Shi et al., 2024). S^2 is parameter-free and enables the extraction of multi-scale features without altering the original pre-trained encoder. By processing sign images at multiple resolutions, S^2 provides a more comprehensive spatial understanding, ensuring that SE captures both fine-grained and broad spatial details for accurate sign language interpretation. The resulting spatial features can be represented as $Z_s \in \mathbb{R}^{T \times 2d}$, where T is the number of frames, and $2d$ is the enhanced embedding dimension, reflecting the integration of multi-scale features.

3.3 Motion Encoder

ME derives motion features from the sign video X. Similar to SE, we employ a pre-trained video encoder (e.g., VideoMAE), which remains frozen, to process sign clips segmented from the video. However, accurately segmenting the sign video into distinct gloss-level clips is challenging without the support of pre-trained Continuous Sign Language Recognition (CSLR) models (Wei and Chen, 2023). To address this limitation, we use a sliding window approach to capture implicit gloss-level representations (Cheng et al., 2023; Hwang et al., 2024). Specifically, we divide the sign video into short, overlapping clips, then feed each clip into ME to extract the implicit gloss-level motion features $Z_m \in \mathbb{R}^{N \times d}$, where $N = \frac{T}{t}$ and t denotes the number of frames between the start of neighboring clips. Since Z_m is generated by processing N short clips, it can also be interpreted as a sequence of N clip-wise features.

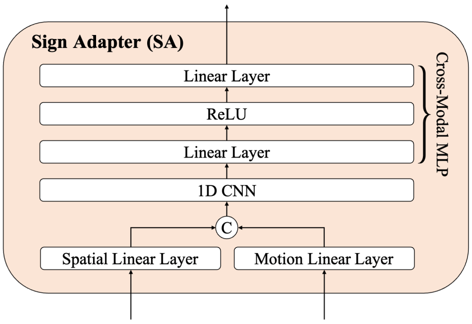


Figure 3: An overview of Sign Adapter.

3.4 Sign Adapter

In the previous sections, we extracted two distinct visual features: the spatial features Z_s and the motion features Z_m. These features differ in both their dimensions and representation, as depicted in Figure 2. To effectively integrate these features, we introduce an additional module called Sign Adaptor (SA). As shown in Figure 3, SA includes linear projection layers, a 1D CNN, and a Multi-Layer Perceptron (MLP). These components work together to integrate the spatial and motion features into a unified sign representation, denoted as Z_{sm}. First, the spatial and motion features are passed through linear projection layers to transform them into features with matching dimensions. Next, the 1D CNN is applied for short-term modeling of the combined features. Finally, a cross-modal MLP (Liu et al., 2024a) is employed to bridge the visual and textual modalities. The resulting outputs are represented as $Z_{sm} \in \mathbb{R}^{M \times d'}$, where M represents the reduced number of frames after convolution, and d' is the dimension aligned with that of the LLM. Although SA aids in bridging the modality gap between visual and textual features during training under the SLT supervision, the gap still persists. To tackle this issue, we introduce VT-Align, which will be detailed in the next section.

3.5 Visual-Text Alignment

VT-Align is a *warm-up and go* process designed to provide the SA module with well-initialized weights before the SLT supervision begins. This initial alignment is crucial, as it helps the model more effectively bridge the modality gap during training. To achieve this alignment, we use a widely-used softmax-based contrastive learning approach (Radford et al., 2021; Jia et al., 2021).

Specifically, given a mini-batch $\mathcal{B} = \{(S_1, Y_1), (S_2, Y_2), ...\}$ of sign-text pairs, the contrastive learning objective encourages the embeddings of matching pairs (S_i, Y_i) to align closely while pushing apart the embeddings of mismatched pairs $(S_i, Y_{j \neq i})$. Text features Z_t are extracted from the target translation Y_i using the LLM's embedding layer $E_{llm}(\cdot)$. Note that only the SA module $f_{sa}(\cdot)$ is updated during this process, while $E_{llm}(\cdot)$ remains fixed to preserve the LLM's language capabilities. The VT-Align loss function L_{vt} is represented as follows:

$$-\frac{1}{2|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\left(\overbrace{\log\frac{e^{\tau Z_{sm}^{(i)} \cdot z_t^{(i)}}}{\sum_{j=1}^{|\mathcal{B}|} e^{\tau Z_{sm}^{(i)} \cdot Z_t^{(j)}}}}^{\text{sign}\rightarrow\text{text softmax}} + \overbrace{\log\frac{e^{\tau Z_{sm}^{(i)} \cdot Z_t^{(i)}}}{\sum_{j=1}^{|\mathcal{B}|} e^{\tau Z_{sm}^{(j)} \cdot Z_t^{(i)}}}}^{\text{text}\rightarrow\text{sign softmax}}\right), \quad (1)$$

where $Z_{sm}^{(i)} = \frac{f_{sa}(S_i)}{\|f_{sa}(S_i)\|_2}$, $Z_t^{(i)} = \frac{E_{llm}(T_i)}{\|E_{llm}(T_i)\|_2}$, and τ denotes a learnable temperature parameter used to scale the logits.

3.6 Training Details

Our framework is optimized in two stages: an initial warm-up phase followed by training with the SLT supervision. In the warm-up phase, we begin by training the SA module using VT-Align for a designated number of steps (e.g., 4K steps). After completing the warm-up phase, we proceed to a joint training of both SA and the LLM. For fine-tuning the LLM, we utilize LoRA (Hu et al., 2021), a lightweight and efficient method specifically designed for this purpose. Overall, our method is trained with a combined loss function as:

$$\mathcal{L}_{SpaMo} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{vt}, \quad (2)$$

where λ is a hyperparameter, and $\mathcal{L}_{ce}$ represents cross-entropy loss.

4 Experiments

4.1 Implementation Details

For SE and ME, we use CLIP ViT-L/14 (Radford et al., 2021) and VideoMAE-L/16 (Tong et al., 2022), respectively. To extract the spatial features, the sign images are interpolated to multiple scales, such as 224×224 and 448×448. For each scale, larger images are split into sub-images of regular size (224×224) and processed individually. These features from the sub-images are then pooled and concatenated with features from the original representation. For the motion features, each clip consists of 16 frames aligned with the findings from (Wilbur, 2009), which suggests that this frame interval captures a single sign. We use an 8-frame gap between neighboring clips. We utilize FlanT5-XL-16bit (Chung et al., 2024) as our LLM. During the warm-up phase with VT-Align, we use 4K steps on PHOENIX14T and 15K steps on How2Sign. Additional implementation details can be found in Appendix Section A.

4.2 Datasets and Evaluation Metrics

Datasets. We evaluated our method on two sign language datasets: PHOENIX14T (Camgoz et al., 2018) and How2Sign (Duarte et al., 2021). **PHOENIX14T** is a German Sign Language (DGS) dataset focused on weather forecasts, featuring a closed domain with a vocabulary of 3K words and an average of 116 frames. **How2Sign** is a large-scale American Sign Language (ASL) dataset that spans over a more open instructional domain, containing a vocabulary of 16K words and an average of 173 frames. Detailed statistics for both datasets are provided in Appendix Section C.

Evaluation Metrics. We report BLEU via SacreBLEU (Papineni et al., 2002; Post, 2018)[2] and

[2] nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1

Setting	Methods	Vis. Ft.	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE
Gloss-based	SLRT (Camgoz et al., 2020)	✓	46.61	33.73	26.19	21.32	-
	STN-SLT (Voskou et al., 2021)	✓	48.61	35.97	28.37	23.65	-
	STMC-T (Zhou et al., 2021b)		46.98	36.09	28.70	23.65	46.65
	BN-TIN-Transf.+SignBT (Zhou et al., 2021a)		50.80	37.75	29.72	24.32	49.54
	SimulSLT (Yin et al., 2021)		48.23	35.59	28.04	23.14	49.23
	PET (Jin et al., 2022)	✓	49.54	37.19	29.30	24.02	49.97
	MMTLB (Chen et al., 2022a)	✓	53.97	41.75	33.84	28.39	52.65
	TS-SLT (Chen et al., 2022b)	✓	54.90	42.43	34.46	28.95	53.48
	SLTUNET (Zhang et al., 2023b)	✓	52.92	41.76	33.99	28.47	52.11
Weakly Gloss-free	TSPNet (Li et al., 2020b)	✓	36.10	23.12	16.88	13.41	34.96
	GASLT (Yin et al., 2023)	✓	39.07	26.74	21.86	15.74	39.86
	ConSLT (Fu et al., 2023)	✓	-	-	-	21.59	47.69
Gloss-free	CSGCR (Zhao et al., 2021)		36.71	25.40	18.86	15.18	38.85
	GFSLT-VLP (Zhou et al., 2023)	✓	43.71	33.18	26.11	21.44	42.29
	FLa-LLM (Chen et al., 2024)	✓	46.29	35.33	28.03	23.09	45.27
	Sign2GPT (Wong et al., 2024)	✓	49.54	35.96	28.83	22.52	**48.90**
	SignLLM (Gong et al., 2024)	✓	45.21	34.78	28.05	23.40	44.49
	SpaMo (Ours)		**49.80**	**37.32**	**29.50**	**24.32**	46.57

Table 1: Performance comparison on the PHOENIX14T dataset. "Vis. Ft." denotes to the visually fine-tuned on sign language datasets. The best results are highlighted in **bold**, and the second-best are underlined.

Setting	Methods	Modality	Vis.Ft.	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	BLEURT
Weakly Gloss-free	GloFE-VN (Lin et al., 2023)	Landmark	✓	14.94	7.27	3.93	2.24	12.61	31.65
	OpenSLT (Tarrés et al., 2023)	RGB	✓	34.01	19.30	12.18	8.03	-	-
Gloss-free	YT-ASL-SLT (Uthus et al., 2024)†	Landmark		14.96	5.11	2.26	1.22	-	29.98
	SSVP-SLT (Rust et al., 2024)†	RGB	✓	30.20	16.70	10.50	7.00	25.70	39.30
	FLa-LLM (Chen et al., 2024)	RGB	✓	29.81	18.99	13.27	9.66	27.81	-
	SpaMo (Ours)	RGB		**33.41**	**20.28**	**13.96**	**10.11**	**30.56**	**42.23**

Table 2: Performance comparison of translation results on the How2Sign dataset. YT-ASL-SLT and SSVP-SLT (marked with †) are reported without dataset scaling to ensure a fair comparison.

ROUGE-L (Lin and Och, 2004). BLEU-n assesses translation precision by evaluating n-grams. ROUGE-L measures text similarity by calculating the F1 score based on the longest common subsequences between predicted and reference texts. We also report BLEURT (Sellam et al., 2020) from the BLEURT-20 checkpoint[3], which has been shown to correlate well with human judgments.

4.3 Comparison with State-of-the-Art

A Taxonomy of SLT. In Section 2, we explore gloss-free methods, including those that incorporate gloss-supervised visual encoders. Although these approaches have traditionally been categorized as gloss-free, we argue that they should more accurately be described as *weakly gloss-free* due to their dependence on gloss-annotated data. This classification is detailed in Table 1. Specifically, methods such as TSPNet (Li et al., 2020b), GASLT (Yin et al., 2023), ConSLT (Fu et al., 2023), GloFE-VN (Lin et al., 2023), and OpenSLT (Tarrés et al., 2023) rely on sign features extracted by visual encoders trained on continuous or isolated sign language recognition (SLR) datasets.

Results on PHOENIX14T. We compare our method with both gloss-based and gloss-free methods on PHOENIX14T. As shown in Table 1, most previous methods rely on the domain-specific fine-tuning of their visual encoders. By contrast, our method demonstrates consistent improvements across all reported metrics without such fine-tuning, except for ROUGE, where it achieves the second-best result. Notably, the improvement on BLEU-4 is particularly significant, with a margin of 0.92, representing a 3.93% increase over SignLLM (Gong et al., 2024). These results demonstrate that conveying the key components of sign language to LLMs is crucial for enhancing translation performance and adopting LLMs in SLT without expensive training or complex training steps.

Results on How2Sign. We also evaluated our method on How2Sign, which poses greater challenges than PHOENIX14T due to its more open-domain nature, longer sign video lengths, and larger vocabulary. The results are presented in Table 2. Our method outperforms previous methods across all reported metrics. Specifically, we achieve a 0.45 margin in BLEU-4, representing a 4.66% improvement over Fla-LLM (Chen et al., 2024).

[3] https://huggingface.co/lucadiliello/BLEURT-20

Component			Metric				
SE	ME	VT-Align	B1	B2	B3	B4	RG
✓			46.44	33.79	26.07	21.11	42.15
	✓		29.71	16.23	10.99	8.36	22.44
✓	✓		47.59	35.05	27.34	22.26	43.92
✓	✓	✓	**49.80**	**37.32**	**29.50**	**24.32**	**46.57**

Table 3: Ablation study of main component.

Models	Params	B1	B2	B3	B4	RG
W/o LLM	60.5M	24.93	12.76	8.45	6.35	18.96
mT5-Large	1.2B	32.31	19.23	13.21	9.87	26.32
Flan-T5-Large	0.8B	47.63	34.75	27.04	22.02	43.66
Flan-T5-XL	3B	**49.80**	**37.32**	**29.50**	**24.32**	**46.57**

Table 4: Ablation study for impact of LLM.

Method	KDEs Entropy ↓
GFSLT-VLP (Zhou et al., 2023)	0.32
SPaMo (Ours)	**0.12**

Table 5: Comparison of KDE entropy values across different embeddings. Lower entropy values indicate more confident and distinct representations.

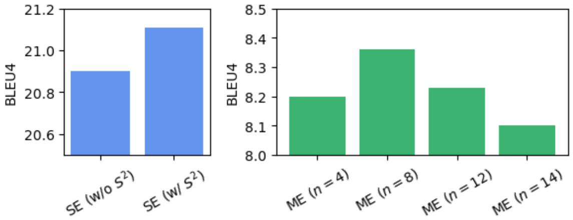


Figure 4: Ablation study for SE and ME. S^2 represents Scaling on Scales, and n denotes the gap between neighboring clips. Note that the results presented do not include VT-Align.

Notably, we see a performance gain in BLEURT, reaching 2.93, which is 7.46% higher than SSVP-SLT (Rust et al., 2024) under the non-scaled dataset setting. These results suggest that our method is more robust for open-domain scenarios with general topics and longer video lengths, even without the domain-specific fine-tuning.

Kernel Density Estimation. To assess the quality of sign representations, following Ye et al. (2023), we employ Kernel Density Estimation (KDE) to estimate the probability density functions of embeddings from GFSLT-VLP and our method on PHOENIX14T. Note that we reproduced GFSLT-VLP using the official code[4]. As shown in Table 5, our method produces more compact and confident representations than GFSLT-VLP. More details can be found in Appendix Section A

4.4 Ablation Study

To further evaluate our method, we conducted extensive ablation experiments on PHOENIX14T, the most widely used sign language dataset. Additional results can be found in Appendix Section B.

Effect of Main Components. We begin by evaluating the effect of the key components in our framework, specifically SE (spatial features), ME (motion features), and VT-align. As shown in Table 3, using both spatial and motion features alone yields results comparable to Sign2GPT in terms of BLEU-4 score (22.52 vs. 22.26). However, when VT-align is incorporated with these features, it achieves the best overall performance, demonstrating the importance of each component in SpaMo.

Effect of LLM. Next, we examine the effect of different LLMs by replacing the model with various LLM types, as shown in Table 4. We compare four baselines, each with a different model size: our method without pre-trained weights, mT5-Large, Flan-T5-Large, and Flan-T5-XL. Of these, Flan-T5-XL demonstrates the best performance.

Effect of S^2 and Neighboring Gap. Finally, we evaluate the effect of S^2 and the gap between neighboring clips on SE and ME, respectively. As shown in Figure 4, S^2 substantially improves translation performance, highlighting its effectiveness to help SE capture more detailed spatial details. Additionally, our analysis reveals that an 8-frame gap between neighboring clips produces the best results, indicating that this specific gap optimally facilitates ME in extracting dynamic motion information.

4.5 Qualitative Analysis

Translation Results. Table 6 presents two example translations on PHOENIX14T, comparing our method with GFSLT-VLP, the only other publicly available baseline. In the first example (top), our method provides an accurate translation, whereas GFSLT-VLP fails to capture the correct semantic meaning. In the second example (bottom), our method again produces a precise translation, while GFSLT-VLP introduces errors, resulting in incorrect information. These examples demonstrate the superior accuracy of our method in generating reliable translations. Please refer to Appendix Section D for more translation examples.

Visual Token Analysis. We performed an additional analysis to explore how the LLM interprets

[4] https://github.com/zhoubenjia/GFSLT-VLP

Ref:	die neue woche beginnt noch wechselhaft und etwas kühler. *(the new week begins still changeable and somewhat cooler)*
GFSLT-VLP:	am montag wieder wechselhaft und kühler. *(on Monday again changeable and cooler)*
Ours:	die neue woche beginnt wechselhaft und wieder kühler. *(the new week begins changeable and again cooler)*
Ref:	sonst viel sonnenschein. *otherwise, a lot of sunshine.*
GFSLT-VLP:	im übrigen land viel sonne. *in the rest of the country, a lot of sun.*
Ours:	sonst viel sonnenschein. *otherwise, a lot of sunshine.*

Table 6: Translation results on the test set compared to GFSLT-VLP on PHOENIX14T. Correctly translated 1-grams are highlighted in blue, while incorrect translations are marked in red.

Vis. Token:	NORDWEST SONST FREUNDLICH STURDY *(NORTHWEST OTHERWISE FRIENDLY STURDY)*
Gloss:	NORDWEST FREUNDLICH *(NORTHWEST FRIENDLY)*
Translation:	richtung norden und westen ist es recht freundlich. *(Towards the north and west it is quite pleasant.)*
Vis. Token:	BLEIBT WIND WINTER *(REMAINS WIND WINTER)*
Gloss:	BLEIBEN WIND *(REMAIN WIND)*
Translation:	es bleibt windig. *(it remains windy.)*
Vis. Token:	LIEBE GUTEN ABEND SCHÖNEN *(DEAR GOOD EVENING BEAUTIFUL)*
Gloss:	GUT ABEND BEGRUESSEN *(GOOD EVENING GREETINGS)*
Translation:	guten abend liebe zuschauer. *(good evening dear viewers.)*

Table 7: Comparison between visual tokens (Vis. Token) and their corresponding glosses. Words highlighted in green are exact matches, those in pink are semantic matches, and words in blue are absent in the gloss but appear in the translation.

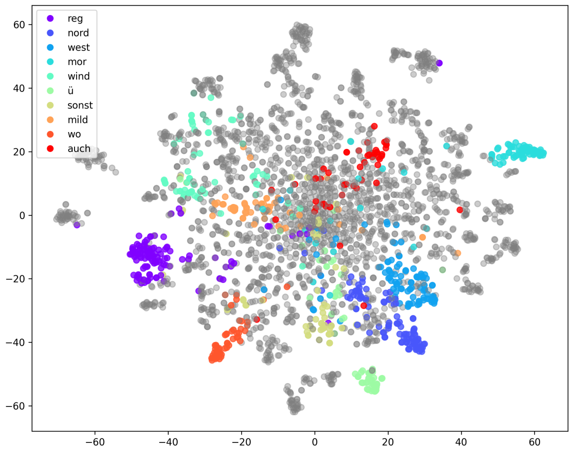


Figure 5: The t-SNE visualization of sign features. Different colors represent sign features with distinct semantics, while gray points are other categories not listed.

the sign videos. Inspired from the reverse engineering (Ju et al., 2023), we first compute the Euclidean distance between the sign feature Z_{sm} and the LLM's embedding table $E_{llm} \in \mathbb{R}^{V \times d'}$, where V represents the vocabulary size. Each sign feature is then mapped to the word associated with the shortest distance in this space. This process can be expressed as $\text{dist}(Z_{sm}, E_{llm}) \leq \Delta$, where $\text{dist}(\cdot)$ denotes the Euclidean distance function, and Δ represents the shortest distance to E_{llm} across all sign features.

Figure 5 shows the t-SNE visualization of each sign feature mapped to the corresponding vocabulary. We observed that certain visual features align closely with specific words, which likely represent the semantic concepts the LLM associates with these features. In other words, these words represent the LLM's interpretation or labeling of the visual content. We refer to these mapped words as “visual tokens”. We further compare these visual tokens with the ground-truth glosses as shown in Table 7. To ensure a clearer and more accurate semantic comparison, repetitive words were removed from the visual tokens. Surprisingly, the LLM's interpretation of the sign videos is similar to the glosses, though not perfectly aligned. This suggests that the LLM has learned to link particular video patterns with specific textual concepts, explaining why those words cluster near the visual features in the embedding space. Additionally, we found that the visual tokens capture words that are present in the translation but not in the glosses. This finding suggests that visual tokens may provide a more comprehensive representation than current glosses, potentially broadening their scope beyond what has been traditionally documented.

5 Conclusion

In this paper, we introduce SpaMo, a novel gloss-free SLT framework based on LLMs. Apart from the previous methods that rely on domain-specific fine-tuning of their visual encoders, SpaMo focuses on capturing the spatial configurations and motion dynamics, eliminating the need for resource-intensive fine-tuning. We also propose VT-Align, a training strategy that effectively aligns and narrows the modality gap between the sign videos and target translations, enabling the transformation of the sign videos into inputs interpretable by the LLM. Our approach achieves state-of-the-art results on two widely-used datasets. Furthermore, we provide the first comprehensive analysis of how the LLM interpret the sign videos within its embedding space and translate them into corresponding text.

Limitations

Limited Use of LLMs. Currently, SpaMo has been tested on a limited range of LLMs. As shown in Table 4, our method scales effectively as the size of the LLMs increases. Therefore, there is a significant opportunity to expand this testing to include various other models, such as Llama (Touvron et al., 2023), Vicuna (Chiang et al., 2023), and Alpaca (Taori et al., 2023). The primary constraint has been the limited availability of GPU resources. Expanding the testing to more LLMs could provide deeper insights and potentially enhance SpaMo's performance across different architectures. Additionally, techniques such as 4-bit quantization can be employed to optimize these models, reducing the computational resources required and enabling more extensive testing. Future work will focus on broadening the range of tested models and exploring these optimization techniques to further improve the scalability and efficiency of SpaMo across diverse LLMs.

Limited Use of Sign Language Datasets. In recent studies (Uthus et al., 2024; Rust et al., 2024), scaling datasets has consistently led to performance improvements, as seen with larger sign language datasets, such as Youtube-ASL (Uthus et al., 2024). While the dataset scaling could also enhance our method, in this work, we focus on a constrained setting. Specifically, we use a limited sign language dataset to evaluate and compare results, demonstrating the effectiveness of our method in resource-limited scenarios. Future work will involve expanding the dataset size to explore the full potential of our method and to assess its scalability and performance across more extensive and diverse datasets.

LoRA Fine-Tuning for LLM. In this paper, we highlight that domain-specific fine-tuning of visual encoders is not essential for our method. However, our method incorporates LoRA fine-tuning for the LLM. While this might appear to be a compromise, it significantly reduces the resource requirements compared to fine-tuning both the visual encoders and the LLM. Additionally, as we discussed in the previous section, this limitation can be addressed as more data becomes available, allowing for improved scalability and performance over time.

Ethics Statement

Our work is focused on developing a practical framework for sign language translation with the goal of overcoming communication barriers faced by the Deaf and hard-of-hearing communities. Although our approach utilizes off-the-shelf visual encoders and LLMs, there is a possibility that the framework could produce unexpected or biased outputs due to the inherent limitations in the pre-trained models. However, we are optimistic that future advancements in LLMs will help mitigate these issues. We rely on open datasets such as PHOENIX14T (Camgoz et al., 2018) and How2Sign (Duarte et al., 2021), which, while containing potentially identifiable information, present minimal concerns regarding personal privacy. Additionally, our method has been validated only on American and German sign languages, limiting its applicability to other sign languages. We call for future research in sign language translation to expand the diversity of sign language datasets, such as YouTube-ASL (Uthus et al., 2024), BOBSL (Albanie et al., 2021), and CSL-Daily (Zhou et al., 2021a), to enhance the framework's applicability and inclusively across different sign languages.

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*.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. 2024. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*.

Rain G Bosworth, Charles E Wright, and Karen R Dobkins. 2019. Analysis of the visual spatiotemporal properties of american sign language. *Vision research*, 164:34–43.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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*, pages 7784–7793.

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*, pages 10023–10033.

Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022a. A simple multi-modality transfer learning baseline for sign language translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5120–5130.

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. 2022b. Two-stream network for sign language recognition and translation. *Advances in Neural Information Processing Systems*, 35:17043–17056.

Zhigang Chen, Benjia Zhou, Jun Li, Jun Wan, Zhen Lei, Ning Jiang, Quan Lu, and Guoqing Zhao. 2024. Factorized learning assisted with large language model for gloss-free sign language translation. In *oceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7071–7081. ELRA and ICCL.

Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, and Wenqiang Zhang. 2023. Cico: Domain-aware sign language retrieval via cross-lingual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19016–19026.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

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.

Karen Emmorey and Shannon Casey. 1995. A comparison of spatial language in english & american sign language. *Sign Language Studies*, 88(1):255–288.

Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J Black. 2024. Chatpose: Chatting about 3d human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2093–2103.

Biao Fu, Peigen Ye, Liang Zhang, Pei Yu, Cong Hu, Xiaodong Shi, and Yidong Chen. 2023. A token-level contrastive framework for sign language translation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Pengzhi Gao, Zhongjun He, Hua Wu, and Haifeng Wang. 2024. Towards boosting many-to-many multilingual machine translation with large language models. *arXiv preprint arXiv:2401.05861*.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. 2023. Continuous sign language recognition with correlation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2529–2539.

Eui Jun Hwang, Sukmin Cho, Huije Lee, Youngwoo Yoon, and Jong C Park. 2024. Universal gloss-level representation for gloss-free sign language translation and production. *arXiv preprint arXiv:2407.02854*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Tao Jin, Zhou Zhao, Meng Zhang, and Xingshan Zeng. 2022. Prior knowledge and memory enriched transformer for sign language translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3766–3775.

Tianjie Ju, Yubin Zheng, Hanyi Wang, Haodong Zhao, and Gongshen Liu. 2023. Is continuous prompt a combination of discrete prompts? towards a novel view for interpreting continuous prompts. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7804–7819.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020a. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1459–1469.

Dongxu Li, Chenchen Xu, Xin Yu, Kaihao Zhang, Benjamin Swift, Hanna Suominen, and Hongdong Li. 2020b. Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. *Advances in Neural Information Processing Systems*, 33:12034–12045.

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

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL-04)*, pages 605–612.

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. 2023. Gloss-free end-to-end sign language translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

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.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 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 Camgöz, and Jean Maillard. 2024. Towards privacy-aware sign language translation at scale. *arXiv preprint arXiv:2402.09611*.

Thibault Sellam, Dipanjan Das, and Ankur P 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.

Baifeng Shi, Ziyang Wu, Maolin Mao, Xin Wang, and Trevor Darrell. 2024. When do we not need larger vision models? *arXiv preprint arXiv:2403.13043*.

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

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

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.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Dave Uthus, Garrett Tanzer, and Manfred Georg. 2024. Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus. *Advances in Neural Information Processing Systems*, 36.

Andreas Voskou, Konstantinos P Panousis, Dimitrios Kosmopoulos, Dimitris N Metaxas, and Sotirios Chatzis. 2021. Stochastic transformer networks with linear competing units: Application to end-to-end sl translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11946–11955.

Fangyun Wei and Yutong Chen. 2023. Improving continuous sign language recognition with cross-lingual signs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23612–23621.

Ronnie B Wilbur. 2009. Effects of varying rate of signing on asl manual signs and nonmanual markers. *Language and speech*, 52(2-3):245–285.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2024. Sign2gpt: Leveraging large language models for gloss-free sign language translation. In *Proceeding of the Eleventh International Conference on Learning Representations*.

Jinhui Ye, Wenxiang Jiao, Xing Wang, Zhaopeng Tu, and Hui Xiong. 2023. Cross-modality data augmentation for end-to-end sign language translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13558–13571.

Aoxiong Yin, Zhou Zhao, Jinglin Liu, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. 2021. Simulslt: End-to-end simultaneous sign language translation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4118–4127.

Aoxiong Yin, Tianyun Zhong, Li Tang, Weike Jin, Tao Jin, and Zhou Zhao. 2023. Gloss attention for gloss-free sign language translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2551–2562.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR.

Biao Zhang, Mathias Müller, and Rico Sennrich. 2023b. SLTUNET: A simple unified model for sign language translation. In *The Eleventh International Conference on Learning Representations*.

Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. 2024a. Motiongpt: Finetuned llms are general-purpose motion generators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 7, pages 7368–7376.

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. 2024b. Why are visually-grounded language models bad at image classification? *arXiv preprint arXiv:2405.18415*.

Jian Zhao, Weizhen Qi, Wengang Zhou, Nan Duan, Ming Zhou, and Houqiang Li. 2021. Conditional sentence generation and cross-modal reranking for sign language translation. *IEEE Transactions on Multimedia*, 24:2662–2672.

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

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021a. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1316–1325.

Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. 2021b. Spatial-temporal multi-cue network for sign language recognition and translation. *IEEE Transactions on Multimedia*, 24:768–779.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

Appendix

In this Supplementary Material, we begin by providing further implementation details in Section A. Section B presents additional experimental results. In Section C, we delve into a more detailed analysis, offering statistics and insights on the sign language datasets used in our study. Finally, in Section D, we conclude by demonstrating additional qualitative results for PHOENIX14T and How2Sign.

A More Implementation Details

Components of SpaMo. In the SA module, we utilize two distinct linear projection layers tailored for the output feature of ME and SE. For short-term modeling, we employ a 1D CNN configured with a specific sequence of layers: $\{K5, P2, K5, P2\}$, where K_σ represents a kernel size of σ, and P_σ indicates a pooling layer with a kernel size of σ (Hu et al., 2023). To integrate features into the LLM's embedding space, we leverage an MLP cross-modal connector (Liu et al., 2024a), projecting the features into a 2048-dimensional space.

Prompt Template. To focus the LLM on the SLT task, we employ a specific prompting strategy. Our prompt includes a clear instructive prompt: "Translate the given sentence into German." Following this, we incorporate multilingual translations via a translation engine such as Google Translator[5], which are sampled from the training set. These translations are included to facilitate In-Context Learning (ICL) (Brown et al., 2020). The prompt is structured as follows: "Translate the given sentence into German. [SRC] = [TRG]." Here, the source input (e.g., a sentence in French) serves as the foreign language example, and the corresponding response is the translation into the target language (e.g., German, as used in PHOENIX14T). An example of this prompt structure is provided in Table 8. To ensure that the LLM does not directly access the target translations during training, we shuffle the translation samples so they do not match the target translation. At test time, we select a translation pair from the training set to use as a reference.

Sign Video Input:	[Extracted Sign Feature]
Instruction:	Translate the given sentence into German.
In Context Examplars:	Soil frost is possible there and in the southern low mountain ranges.=dort sowie in den südlichen mittelgebirgen ist bodenfrost möglich. La helada del suelo es posible allí y en las cadenas montañosas del sur.=dort sowie in den südlichen mittelgebirgen ist bodenfrost möglich. Le gel du sol est possible là-bas et dans les chaînes de montagnes basses du sud.=dort sowie in den südlichen mittelgebirgen ist bodenfrost möglich.

Table 8: An example of prompt used in this paper.

Visual Encoders (SE + ME)	**B1**	**B2**	**B3**	**B4**	**RG**
DINOv2 + V-JEPA	45.67	32.94	25.27	20.35	41.32
DINOv2 + VideoMAE	47.31	34.60	26.90	21.86	42.50
CLIP + V-JEPA	47.82	34.71	26.76	21.66	43.68
CLIP + VideoMAE	**49.80**	**37.32**	**29.50**	**24.32**	**46.57**

Table 9: Ablation study on various combinations of visual encoders. The results are with VT-Align.

Methods	**B1**	**B2**	**B3**	**B4**	**RG**
Ours (w/o LoRA)	46.11	32.65	24.69	19.67	42.91
Ours (w/ LoRA)	**49.80**	**37.32**	**29.50**	**24.32**	**46.57**

Table 10: Ablation on our method with and without LoRA.

Training. For training, we use the AdamW optimizer (Loshchilov and Hutter, 2017), with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and a weight decay of 0.01. The learning rate schedule includes a cosine decay with a peak learning rate of 1e-4 and a linear warmup of than 10k steps, with a minimum learning rate of 5e-5. We train our model for 40 epochs, using a single NVIDIA A100 GPU, completing the entire process within 24 hours.

Evaluating Process with KDEs. To evaluate the quality of the learned representations, we utilize Kernel Density Estimation (KDE) to estimate the probability density functions of the embeddings from GFSLT-VLP and ours. Due to different dimensionality from these methods (1,024 vs. 2,048), we run Principal Component Analysis (PCA) to reduce the number of dimensions while retaining the most significant variance components. This dimensionality reduction facilitated more efficient and stable KDE fitting. KDE can be expressed as:

$$f_{\text{kde}}(\mathbf{z}) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{z} - \mathbf{z_i}}{h}\right), \qquad (3)$$

where $\mathbf{z_i}$ denotes the representation points, K denotes the kernel function, h is the bandwidth parameter, d is the dimensionality of the data, and n is the number of data points.

The entropy of KDE is then calculated as:

$$H = -\sum_{i=1}^{n} f_{\text{kde}}(\mathbf{z_i}) \log f_{\text{kde}}(\mathbf{z_i}), \qquad (4)$$

[5] https://cloud.google.com/translate

Dataset	Language	# Vocab	Train / Valid / Test	Avg. No. Frame	Gloss	Domain
PHOENIX14T (Camgoz et al., 2018)	DGS	3K	7,096 / 519 / 642	116	O	Weather Forecast
How2Sign (Duarte et al., 2021)	ASL	16K	31,128 / 1,741 / 2,322	173	X	Instructional

Table 11: Statistics of two sign language datasets used in this work. DGS: German Sign Language; ASL: American Sign Language; Avg. No. Frame: average number of video frames.

where H represents the entropy, and $f(\mathbf{z_i})$ are the estimated density values at the representation points.

B More Experiments

Effect of Visual Encoders. We assess the effect of various combination of visual encoders (SE & ME). Table 9 shows four different encoders: DINOv2 (Oquab et al., 2023), CLIP (Radford et al., 2021), V-JEPA (Bardes et al., 2024), and VideoMAE (Tong et al., 2022). The results demonstrate that the combination of CLIP and VideoMAE delivers the highest performance, suggesting potential for further improvement as visual encoders continue to advance.

Effect of LoRA. We evaluate the effect of LoRA on the LLM. As illustrated in Table 10, the LLM with LoRA demonstrates superior performance.

C Statistics of Sign Language Datasets

Table 11 provides an overview of two sign language datasets: PHOENIX14T and How2Sign, which differ significantly in their characteristics and applications. PHOENIX14T focuses on German Sign Language (DGS) within the specific domain of weather forecasting, featuring a relatively small vocabulary of 3K words and a concise average video length of 116 frames. It includes 7,096 training samples, 519 validation samples, and 642 test samples, with gloss annotations available. This dataset is tailored for domain-specific tasks, offering clear and repetitive patterns ideal for translation and recognition within weather-related contexts.

In contrast, How2Sign, centered on American Sign Language (ASL) in the instructional domain, presents a much larger and more diverse dataset with a 16K word vocabulary and an average of 173 frames per video. It provides a substantial amount of data, with 31,128 training samples, 1,741 validation samples, and 2,322 test samples, though it lacks gloss annotations. The broader and more complex nature of How2Sign makes it suitable for general sign language processing tasks, especially those requiring an understanding of diverse and intricate sign sequences.

D More Qualitative Results

We provide additional translation examples for PHOENIX14T and How2Sign in Tables 12 and 13, respectively. As shown in Table 12, our method consistently delivers accurate translations, while GFSLT-VLP struggles to capture the correct semantic meaning.

For How2Sign, Table 13 presents translation results along with their corresponding visual tokens. Since How2Sign lacks gloss annotations, we include actual sign frames for qualitative comparison. Similar to the PHOENIX14T results, several visual tokens in How2Sign are closely aligned with the translations. Note that although OpenSLT (Tarrés et al., 2023) is the only publicly available baseline[6], we were unable to reproduce their results due to a broken link to the fine-tuned I3D features at the time of drafting.

[6] https://github.com/imatge-upc/slt_how2sign_wicv2023

Ref:	und nun die wettervorhersage für morgen sonntag den zwölften juli. *(and now the weather forecast for tomorrow Sunday the twelfth of July.)*
GFSLT-VLP:	und nun die wettervorhersage für morgen sonntag den zwölften juni. *(and now the weather forecast for tomorrow, Sunday, the twelfth of June.)*
Ours:	und nun die wettervorhersage für morgen sonntag den zwölften juli. *(and now the weather forecast for tomorrow Sunday the twelfth of July.)*
Ref:	in der nacht muss vor allem in der nordwesthälfte mit schauern und gewittern gerechnet werden die heftig ausfallen können. *(During the night, showers and thunderstorms are expected, especially in the northwest half, which could be heavy.)*
GFSLT-VLP:	heute nacht gibt es im norden teilweise kräftige schauer und gewitter die örtlich unwetterartig sein können. *(At night, showers and thunderstorms can be expected, especially in the northwest half, which can sometimes be strong.)*
Ours:	in der nacht muss vor allem in der nordwesthälfte mit schauern und gewittern gerechnet werden die mitunter kräftig sein können. *(During the night, showers and thunderstorms are expected, particularly in the northwest half, which may be heavy.)*
Ref:	und nun die wettervorhersage für morgen donnerstag den siebenundzwanzigsten august. *(and now the weather forecast for tomorrow, Thursday the twenty-seventh of August.)*
GFSLT-VLP:	und nun die wettervorhersage für morgen donnerstag den sechsundzwanzigsten august. *(and now the weather forecast for tomorrow, Thursday the twenty-sixth of August.)*
Ours:	und nun die wettervorhersage für morgen donnerstag den siebenundzwanzigsten august. *(and now the weather forecast for tomorrow, Thursday the twenty-seventh of August.)*
Ref:	am tag ist es im westen freundlich sonst sonne und dichtere wolken im wechsel hier und da fallen einzelne schauer. *(During the day it is friendly in the west, otherwise sun and denser clouds alternate, with occasional showers here and there)*
GFSLT-VLP:	am tag wechseln sonne und wolken einander ab im westen fallen mitunter gewittrige schauer. *(During the day sun and clouds alternate, in the west, occasional stormy showers may occur)*
Ours:	am tag ist es im westen freundlich mit sonne und dichteren wolken hier und da fallen schauer. *(During the day it is friendly in the west with sun and denser clouds, with occasional showers here and there)*
Ref:	abseits der gewittern weht der wind schwach bis mäßig an der küste frisch. *(Away from the thunderstorms, the wind blows weak to moderate, fresh at the coast.)*
GFSLT-VLP:	abgesehen von gewitterböen schwacher bis mäßiger an den küsten auch frischer wind *(Apart from thunderstorm gusts, weak to moderate, also fresh wind at the coasts.)*
Ours:	abseits der gewittern weht der wind schwach bis mäßig an den küsten auch frisch. *(Away from the thunderstorms, the wind blows weak to moderate, also fresh at the coasts.)*
Ref:	am sonntag im norden und an den alpen mal sonne mal wolken und ab und an schauer sonst ist es recht freundlich. *(On Sunday in the north and in the Alps sometimes sun sometimes clouds and occasional showers otherwise it is quite pleasant.)*
GFSLT-VLP:	am sonntag im norden an den alpen einige schauer sonst ist es recht freundlich. *(On Sunday in the north in the Alps some showers otherwise it is quite pleasant.)*
Ours:	am sonntag im norden und an den alpen mal sonne mal wolken und nur einzelne schauer sonst meist freundlich. *(On Sunday in the north and in the Alps sometimes sun sometimes clouds and only a few showers otherwise mostly pleasant.)*
Ref:	am mittwoch eine mischung aus sonne wolken und nebelfeldern im nordwesten hier und da schauer sonst ist es trocken. *(On Wednesday a mix of sun, clouds, and fog patches in the northwest; here and there showers, otherwise it is dry.)*
GFSLT-VLP:	am mittwoch gibt es viele wolken hier und da schauer vor allem im nordwesten bleibt es meist trocken. *(On Wednesday there will be many clouds; here and there showers, especially in the northwest, it remains mostly dry.)*
Ours:	am mittwoch eine mischung aus sonne wolken und nebel im nordwesten einige schauer sonst bleibt es meist trocken. *(On Wednesday a mix of sun, clouds, and fog in the northwest; some showers, otherwise it remains mostly dry.)*
Ref:	am tag scheint verbreitet die sonne im süden und westen bilden sich später gebietsweise quellwolken. *(During the day, the sun shines widely in the south, and later, isolated cumulus clouds form in the west.)*
GFSLT-VLP:	am tag scheint in der südhälfte häufig die sonne hier und da ein paar wolken. *(During the day, the sun often shines in the southern half, here and there a few clouds.)*
Ours:	am tag scheint verbreitet die sonne im süden und im äußersten westen tauchen hier und da ein paar quellwolken auf. *(During the day, the sun shines widely in the south, and in the far west, here and there, a few cumulus clouds appear.)*
Ref:	der wind weht mäßig bis frisch mit starken bis stürmischen böen im bergland teilweise schwere sturmböen im südosten mitunter nur schwacher wind. *(The wind blows moderately to freshly with strong to stormy gusts in the mountainous regions, partly severe storm gusts in the southeast, occasionally only weak wind.)*
GFSLT-VLP:	der wind weht mäßig bis frisch bei schauern sowie im südosten schwere sturmböen im bergland starker bis stürmböen. *(The wind blows moderately to freshly with showers, as well as severe storm gusts in the southeast, in the mountainous regions strong to stormy gusts.)*
Ours:	der wind weht mäßig bis frisch mit starken bis stürmischen böen auf den bergen schwere sturmböen im süden sonst schwacher wind. *(The wind blows moderately to freshly with strong to stormy gusts on the mountains, severe storm gusts in the south, otherwise weak wind.)*
Ref:	am montag überall wechselhaft und deutlich kühler. *(On Monday, everywhere is changeable and significantly cooler.)*
GFSLT-VLP:	am montag wird es wieder wechselhafter kühler. *(On Monday, it will be changeable and cooler again.)*
Ours:	am montag überall wechselhaft und deutlich kühler. *(On Monday, everywhere is changeable and significantly cooler.)*
Ref:	sonst ein wechsel aus sonne und wolken. *(Otherwise a mix of sun and clouds.)*
GFSLT-VLP:	ansonsten wechseln sich teilweise dichte wolken und sonne ab. *(Otherwise partially dense clouds and sun alternate.)*
Ours:	sonst ein wechsel aus sonne und wolken. *(Otherwise a mix of sun and clouds.)*
Ref:	und nun die wettervorhersage für morgen samstag den sechsundzwanzigsten januar. *And now the weather forecast for tomorrow, Saturday, the twenty-sixth of January.*
GFSLT-VLP:	und nun die wettervorhersage für morgen samstag den sechsundzwanzigsten dezember. *And now the weather forecast for tomorrow, Saturday, the twenty-sixth of December.*
Ours:	und nun die wettervorhersage für morgen samstag den sechsundzwanzigsten januar. *And now the weather forecast for tomorrow, Saturday, the twenty-sixth of January.*
Ref:	sonst ist es recht freundlich. *Otherwise it is quite pleasant.*
GFSLT-VLP:	sonst überwiegend freundlich. *Otherwise mostly pleasant.*
Ours:	sonst ist es recht freundlich. *Otherwise it is quite pleasant.*

Table 12: Translation results on the test set compared to GFSLT-VLP on PHOENIX14T. Correctly translated 1-grams are highlighted in blue, while incorrect translations are marked in red.

Image:	
Vis. Token:	AGAIN SOMEONE ONE SHOW
Ref:	again, one more time we'll show it for you.
Ours:	again, one more time.
Image:	
Vis. Token:	LITTLE MORE HOW
Ref:	a little bit more then this maybe.
Ours:	a little bit more about it.
Image:	
Vis. Token:	NOW GO TODAY TO TAKE LITTLE THREE SEVEN FOUR WEED OUT LITTLE HERE JUILLET VORSCHRIFTEN
Ref:	and we're going to take a little weed out here.
Ours:	now we're going to take a little bit of the weed out here.
Image:	
Vis. Token:	WANT TO REPEAT TWO LOOK ON YOURÄNG KISS AGE IS YOUR HORSE
Ref:	you want to look at the age of your horse.
Ours:	you want to take a look at the age of your horse.
Image:	
Vis. Token:	MANY PEOPLE NOT OTHER UNDERSTAND THOUGHT
Ref:	many people don't understand.
Ours:	many people don't understand that.
Image:	
Vis. Token:	I PRACTICE WHEN WITH B FOAMERS CAST WAS SO OROU CAN KNOW IF OR GROUP
Ref:	i practice with the barton oaks dental group.
Ours:	i practice with the barton oaks tennis team.
Ref:	so, let's keep doing the same thing with the arms.
Ours:	so, let's keep doing the same thing with the arms.
Ref:	here, two, three, four, elbow and follow wherever you're going to go, like the knee to the groin and your elbow.
Ours:	here, two, three, four, follow through where you're going to want to squeegee, woo, woo, your elbow.
Ref:	my name is robert segundo and have fun.
Ours:	my name is robert todd and have fun.
Ref:	watch our next segment to learn more about natural beauty products.
Ours:	watch our next segment and we'll talk a little bit more about natural beauty products.
Ref:	remember, be careful when doing your home remedies, and if you're not sure, check with your local professional.
Ours:	remember very carefully when doing your home remedies if you have a cell phone.
Ref:	you can start to rotate your shoulders and start to get more comfortable with your feet by turning.
Ours:	you can start rotating your shoulders and start getting comfortable with your five by rotating.
Ref:	hi, i'm johanna krynytzky with hip expressions belly dance studio in st. petersburg, florida.
Ours:	hi, i'm johanna krynytzky with hip expressions belly dance studio in st. petersburg, florida.
Ref:	i'm going to show you how to do some step-touch side foot work for belly dancing.
Ours:	i'm going to show you some step touch side and medium rock for belly dancing.

Table 13: Translation results on the How2Sign test set. Correctly translated 1-gram matches are highlighted in blue. Exact visual token matches within the translation are highlighted in green .