ITC 1/55
Information Technology and Control
Vol. 55 / No. 1/ 2026
pp. 188-203
DOI 10.5755/j01.itc.55.1.40498

Multimodal Large Language Model for Gloss-Free Video Sign Language Translation

Received 2025/02/13 | Accepted after revision 2025/06/26

HOW TO CITE: Guo, R., Hu, X., Peng, T., Du, Y. (2026). Multimodal Large Language Model for Gloss-Free Video Sign Language Translation. *Information Technology and Control*, 55(1), 188-203. https://doi.org/10.5755/j01.itc.55.1.40498

Multimodal Large Language Model for Gloss-Free Video Sign Language Translation

Rong Guo
School of Biological Science and Medical Engineering, Ministry of Education, Beijing Advanced Innovation Center for Biomedical Engineering, Beihang University, Beijing, China; e-mail: guorong@buaa.edu.cn

Xiaohui Hu
Science and Technology on Integrated Information System Laboratory Institute of Software Beijing, 100190, China; e-mail: hxh@iscas.ac.cn

Taiying Peng, Yao Du*
School of Automation Science and Electrical Engineering, Beihang University; Beijing 100191, China; e-mail: duyaoo@buaa.edu.cn

Corresponding author: duyaoo@buaa.edu.cn

The pre-trained large language models (LLMs) achieve impressive advancements not only in text-based tasks but also show significant potential in basic visual-language comprehension. However, it remains uncertain whether LLMs pre-trained on text can comprehend the grammar of sign language based on gestural actions after fine-tuning with limited data. Despite the near-human performance of current LLMs in understanding and generating spoken text, their ability to transition from text language to visual language for multimodal tasks is still confusing. In this paper, we propose the SL-LLaMA model, which leverages the robust capabilities of LLMs for sign language translation tasks and investigate the multimodal abilities of LLMs in transferring from textual language to visual language. We use the LLaMA 2 family of models to perceive and understand sign language grammar and to generate corresponding spoken text. To incorporate video information into the LLM, we propose a sign language translation framework that integrates a vision encoder, an MM-Adaptor, and an LLM to understand sign language and generate the spoken language. Additionally, we employed the *language alignment-Supervised fine-tuning* training strategy to infuse sign language knowledge into the model. Our study evaluates the performance of gloss-free sign language translation on two benchmarks: RWTH-PHOENIX-Weather-2014-T and CSL-Daily. Compared to current state-of-the-art methods, the proposed model achieves competitive results, demonstrating the strong potential of text-pretrained LLMs in understanding visual gram-

matical knowledge. Ablation experiments explore the impact of each component on sign language translation, as well as the framework's generalization and scalability, providing a foundational basis for future applications of LLMs in more complex multimodal tasks.

KEYWORDS: Large Language Model, Sign language translation, Multi-modal, Low-Rank Adaptation.

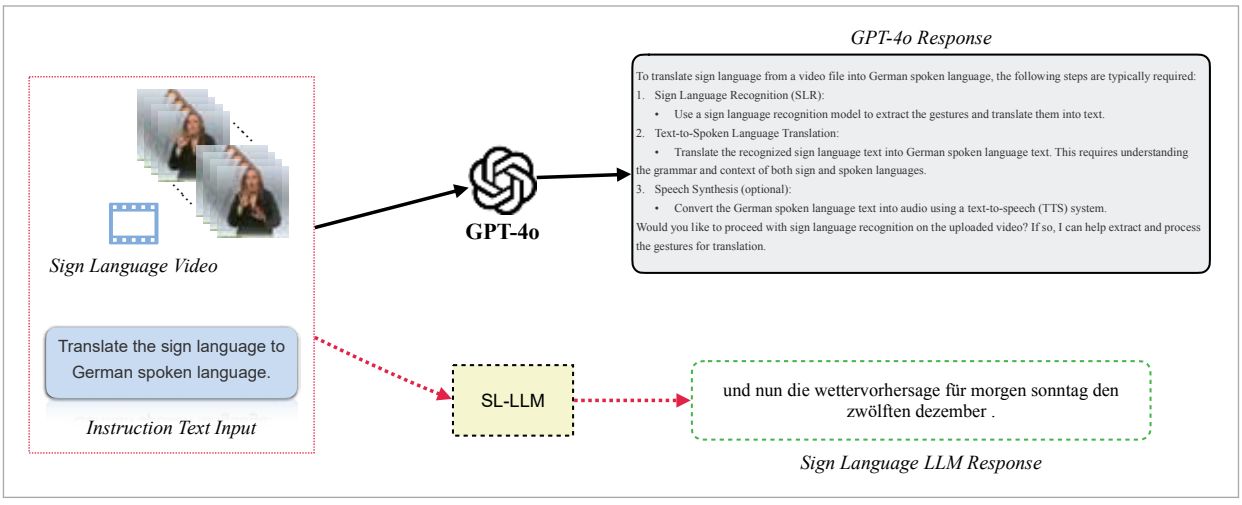
1. Introduction

Sign language is a crucial communication method for individuals with hearing impairments, with millions worldwide relying on it for daily interactions [3, 5]. However, the inherent complexities of sign language—its unique grammatical structures, reliance on multimodal cues (e.g., hand movements, facial expressions, and body postures), and its vast divergence from spoken languages—make developing effective computational translation methods a longstanding challenge [41]. Gloss-free sign language translation (SLT) has emerged as a promising research area, eliminating the need for intermediate linguistic annotations, and thus simplifying the data annotation process [40]. Despite progress, the field still grapples with critical issues: accurately aligning multimodal visual signals to textual outputs, comprehending sign language grammar, and overcoming the scarcity of high-quality datasets [10].

Recent advancements in Large Language Models (LLMs) demonstrated unprecedented capabilities in understanding and generating natural languages, as well as excelling in cross-modal tasks such as image caption. Multimodal large language models (MLLMs) like GPT-4o [14] and LLaVA [19] showcase the potential of LLMs in integrating visual and linguistic information. Nevertheless, their application to sign language translation remains largely underexplored. Video-based tasks like SLT require not only static but also the ability to process sequential visual frames and derive nuanced grammatical representations from gestures-the difficulty beyond the capabilities of existing multimodal LLMs optimized for image-text tasks, as shown in Figure 1. On the one hand, sign language translation introduces unique challenges, including dynamic visual semantic understanding and the representation of motion-based grammatical constructs. Current MLLMs primarily focus on static image-text pairs, often neglecting the temporal and grammatical complexities inherent in video tasks. In addition, training large models typically requires

Figure 1

Sign language translation in MLLM. The current advanced multimodal large language model, GPT-4o, is unable to effectively perform sign language translation. Instead, it is limited to answering queries related to the steps of the translation process. Our MLLM is designed to follow explicit instructions and directly generates the translated text from sign language.



large-scale data, whereas annotated sign language data is highly limited. Therefore, the critical challenge in achieving SLT by MLLM is how to design a reasonable model structure and implement effective model fine-tuning strategy under data constraints.

To address the challenges, we design a multimodal large language model framework for gloss-free SLT and introduce SL-LLaMA, alone with a proposed fine-tuning strategy. SL-LLaMA incorporates a vision encoder, a specialized MM-Adaptor and a pretrained LLM backbone. We adopt the two-stage strategy, *Language Alignment-Supervised Fine-tuning*, to train the model, endowing it with sign language translation capabilities. To avoid overfitting and catastrophic forgetting with limited annotated data, inspired by research [12], we utilize the Low-Rank Adaptor (LoRA) method to fine-tune the LLM parameters. The primary contributions of this paper are as follows:

1 We develop SL-LLaMA, a cross-modal sign language translation framework combining a pretrained LLM, a vision encoder, and an MM-Adaptor to bridge visual and textual modalities.

2 We introduce the Alignment-SFT approach, enabling efficient fine-tuning of LLMs with LoRA to adapt to the complex grammar of sign language using limited data.

3 Experimental validation on two benchmarks illustrates the effectiveness of SL-LLaMA in gloss-free translation tasks, showcasing its scalability to broader vision-language challenges. Generalization experiments also demonstrate the excellent adaptability and scalability.

2. Related Work

2.1. Sign Language Translation

Sign language translation (SLT) aims to translate continuous sign language videos into corresponding spoken language sentences. As it involves both video and text modalities, research and improvements in this field have predominantly focused on two key aspects: visual feature extraction and temporal sequence modeling [38].

In terms of visual feature extraction, most SLT methods employ convolutional neural networks to extract spatial features from individual frames of sign language videos [7, 16, 33]. With advancements in computer vision technologies, more effective visual backbone networks, such as ResNets [25] and EfficientNets [5], have been utilized for capturing sign language movements, thereby improving translation quality. Recently, a limited number of studies explored the use of Vision Transformers (ViTs) for feature extraction [9]. This approach effectively captures global visual information, enabling the extraction of semantic-rich features that are beneficial for understanding the meaning of sign language gestures.

For temporal sequence modeling, SLT has drawn extensively from methods in text-based machine translation [22]. Early approaches, such as Hidden Markov Models (HMMs) [15, 16], have evolved into the use of RNNs, with LSTM networks being widely adopted [4]. Similar to visual technologies, advancements in machine translation have also propelled progress in sign language translation, have been increasingly applied in recent years, significantly enhancing translation performance [34]. Recent SLT methods have increasingly adopted encoder-decoder frameworks and Transformers to model complex sign language structures. Notable works in this direction include TSPNet [18], SLTR-T [5], and GASLT [36], which utilize hierarchical visual encoders and sequential decoders. Signformer [35] and GFSLT [40] further advanced this field by introducing attention-based models and leveraging visual-language pretraining.

Large language models (LLMs) have demonstrated remarkable capabilities in recent time [1, 29]. Through scaling laws [11] and emergent properties [31], LLMs exhibit advanced abilities such as logical reasoning, writing, and role-playing, which are not achievable in smaller models [24, 30]. However, the application of LLMs in sign language translation remains underexplored. Sign2GPT integrates a small-scale LLM with a vision encoder to perform gloss-free SLT [32]. It proposes a pseudo-gloss pretraining strategy to enhance visual-text alignment. While Sign2GPT demonstrates strong performance, its task-specific tuning limits its generalizability to broader multimodal scenarios.

Although some exploratory studies have been conducted [32], models exceeding 7 billion parameters have yet to be utilized for sign language translation tasks. This highlights a promising avenue for lever-

aging the power of LLMs to further advance sign language translation research and development.

2.2. LLM for Vision-Text Tasks

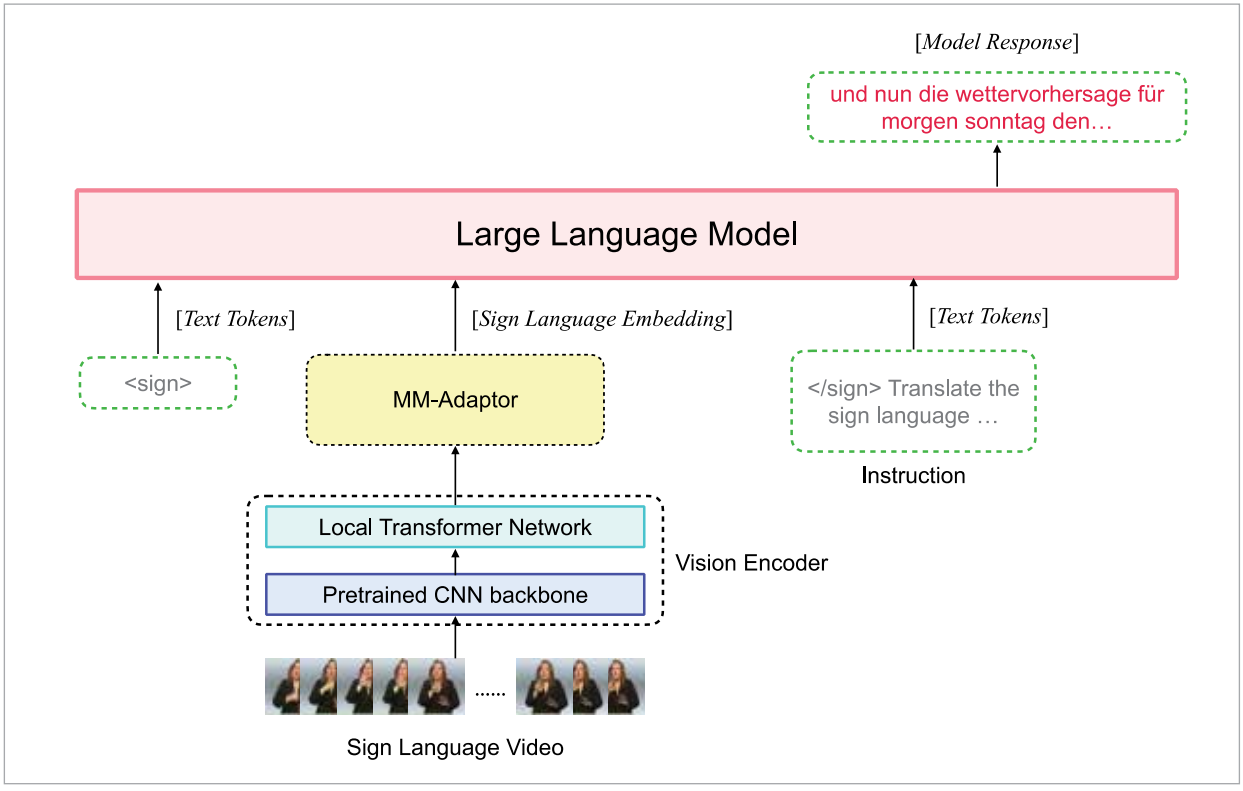
Large language models (LLMs) have demonstrated groundbreaking achievements in image-text tasks, such as image captioning, visual question answering (VQA), and mathematical reasoning. Models like MiniGPT-4 [42] and LLaVA [19] have successfully leveraged the extensive pre-trained knowledge of LLMs, showcasing their capability to effectively understand and process both visual and textual inputs.

Generally, these models integrate a vision encoder and an LLM. MiniGPT-4 [42] employs CLIP-ViT as a vision encoder and a Q-Former with a frozen LLM for image captioning. LLaVA [19] incorporates a visual instruction tuning pipeline with LLaMA and Vicuna as the text decoder. Visual networks such as CLIP-ViT [27], Dinov2 [23], and SigLip [37] are used to process images into a series of patch embeddings, which are then passed through an adaptor before being fed into the LLM. The choice of adaptor varies across models: MiniGPT-4 employs a simple linear mapping layer to convert image feature vectors into the dimensional space of word embeddings; LLaVA utilizes a two-layer fully connected network; and MM1 [21] and Qwen-VL [2] adopt attention-based networks as adaptors. These modules, once trained, enable the LLM to effectively utilize the visual information provided by the vision encoder.

Despite the rapid progress of LLMs in image-text tasks, their performance degrades significantly in video-based tasks due to the lack of temporal modeling and insufficient alignment with dynamic visual semantics. However, the success of LLMs in image-text tasks inspires further exploration of their potential in this domain.

Figure 2

The architecture of SL-LLaMA. The model consists of three main components: the Vision Encoder, the MM-Adaptor, and the LLM backbone. The Vision Encoder includes a pre-trained CNN and a lightweight Local Transformer network. The MM-Adaptor is composed of two linear layers with a nonlinear activation function. The LLM backbone is based on the LLaMA family of models and is primarily responsible for understanding instructions and sign language semantics, and generating the translated text.



In the context of sign language translation, LLMs offer unique advantages. Their pre-trained linguistic capabilities empower them to generate coherent and contextually appropriate textual outputs, addressing the challenge of semantic alignment between video frames and corresponding textual descriptions. This positions LLMs as a promising avenue for advancing the field of sign language translation, leveraging their robust capabilities to bridge the gap between visual and textual modalities.

3. Methods

3.1. Model Structure

Our proposed SL-LLaMA model, as illustrated in Figure 2, comprises three components: a Vision Encoder, an MM-Adaptor, and a large language model backbone. Each module is selected based on their complementary roles in handling the unique challenges of sign language translation. First, the Vision Encoder processes sequential sign language video frames, which are rich in spatial and temporal information. The MM-Adaptor is introduced to bridge the visual features and the LLM's textual input space because of the gap on the two modalities. The LLM is selected due to its strong reasoning and language generation ability.

Vision Encoder. In mainstream multimodal large language models, such as LLaVA, pre-trained CLIP-ViT models are employed as vision encoder to process image signals. In our SL-LLaMA, we first utilize a pre-trained CNN network to process each frame of a sign language video.

$$Z = \{z_1, z_2, \ldots, z_T\}, z_i = f_{CNN}(v_i) \quad (1)$$

Specifically, for a given sign language video $V = \{v_1, v_2, \ldots, v_T\}, v_i \in R^{T\times 3\times 224\times 224}$, the spatial feature of each frame is extracted as $Z = \in R^{T\times 1024}$.

$$L = f_{LTN}(Z). \quad (2)$$

Since sign language conveys semantic information through an ordered sequence of frames, we design a local Transformer network (LTN) within the vision encoder to learn the spatiotemporal feature representation $L \in R^{T\times 1024}$, which provides the SL-LLaMA with semantic information from sign language. Specifically, for each input token corresponding to a frame, the attention is computed only within a fixed-size temporal window (e.g., ±4 frames when the window size is 8). Tokens outside this window are masked, and their attention weights are set to negative infinity before softmax. The window size of the local self-attention is set to 8, based on our analysis showing that the average ratio of video length to text length in sign language datasets is approximately 8. In summary, the vision encoder processes sign language videos through a combination of CNN and self-attention networks, encoding visual data into spatiotemporal feature representations to effectively capture the semantic content of the input.

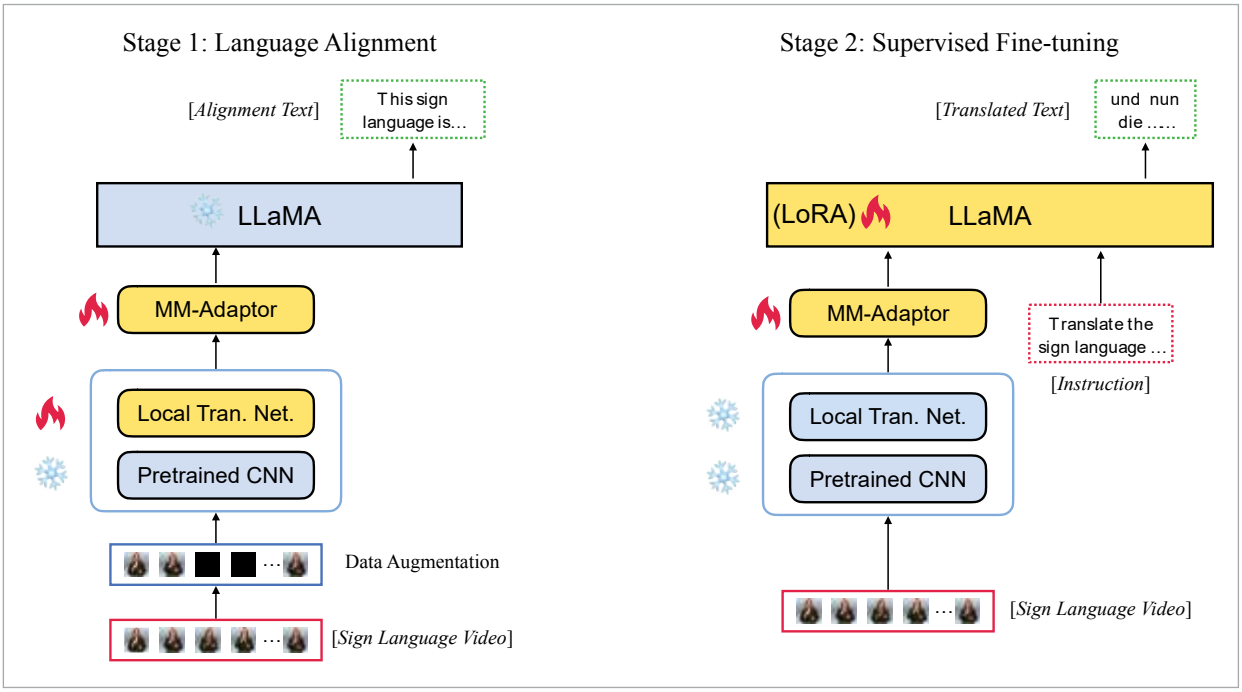
MM-Adaptor. To enable the input of video features processed by the vision encoder into the LLM for modeling, a dedicated module is required to handle these features. Commonly used modules include linear layers [29], cross-attention mechanisms [2], and C-Abstractors [21]. We employed an MLP with a nonlinear activation function. It serves as a bridge connecting linguistic and visual information, processing the spatiotemporal representation of sign language generated by the vision encoder. Our MM-Adaptor consists of two linear layers and a GELU activation function, mapping the vectors L to a 4096-dimensional space as $S \in R^{T\times 4096}$ before injecting them into the LLM.

$$S = f_{GLEU}(L * W_1) * W_2, \quad (3)$$

where $W_1 \in R^{1024\times 4096}$ and $W_2 \in R^{4096\times 4096}$ are the weights of the linear layers. Compared to linear layers, we analyzed that MLPs are more suitable due to the necessity of the nonlinear activation function for the model. Our ablation experiments support this analysis.

LLM. In SL-LLaMA, we utilize the LLaMA family [30] of models as the LLM backbone for perceiving, understanding, reasoning, and generating multimodal information from visual-text inputs. The LLaMA model is an open-source, text-pretrained LLM that supports multiple languages. Many recently proposed LLMs [8] and MLLMs [42] are derived from the LLaMA model through further pre-training and fine-tuning. We employ the LLaMA model to generate natural and fluent spoken

Figure 3
The training pipeline of the SL-LLaMA model. We design two-stage training strategy. In language alignment, we utilize the "sign language - spoken text" data pairs to train the parameters of local Transformer network in vision encoder and MM-Adaptor. During the SFT stage, the sign language and the instruction are the input of model. The MM-Adaptor and LLM backbone are trainable, while the LLM is fine-tuned with LoRA.



language based on the input sign language video features. The LLM takes visual features and textual instruction tokens as input. The tokens are processed into word embeddings by tokenizer, and the LLM then autoregressively generates the translated spoken language text following the instructions. It is important to note that LLMs like LLaMA are pretrained exclusively on textual corpus and, therefore, inherently lack the ability to understand sign language videos. To address this imitation, we design the training strategy to align the sign language and the text for SL-LLaMA.

3.2. Training Strategy

To train SL-LLaMA to learn sign language representation and generation using a limited training set, we adopt a two-stage fine-tuning method called "Language Alignment-Supervised Fine-tuning (SFT)", as shown in Figure 3. We intend to explore the ability of a large language model to extend to multimodal capabilities without any prior visual-text data pre-training. Therefore, we train and fine-tune the SL-LLaMA using only the training set of the sign language dataset, without incorporating any other large-scale video-text data.

Language Alignment. In the alignment learning stage, we aim to inject grammatical and semantic knowledge of sign language into the model by aligning visual features with spoken text. We construct training pairs consisting of sign language videos and their corresponding spoken language sentences. The Local Transformer Network (LTN) in the Vision Encoder and the MM-Adaptor are set as trainable, while the LLM remains frozen at this stage.

The visual input for each video is processed into a spatiotemporal sequence, which is then projected to a 4096-dimensional embedding by the MM-Adaptor. This aligned feature sequence serves as input tokens for the LLM. During training, we apply a mild data augmentation strategy by randomly dropping 10% of video frames. This improves robustness while preserving overall semantics due to the high frame redundancy in sign language videos.

The alignment loss is computed as the cross-entropy between the generated text and ground-truth sentences, enabling the visual modules to learn text-aligned semantics and temporal dependencies in gesture sequences.

Supervised Fine-Tuning. In the SFT stage, we fine-tune both the MM-Adaptor and the LLM using the annotated training set. There are two primary objectives in the supervised fine-tuning stage. First, it aims to enhance the semantic understanding of SL videos and strengthen the alignment between visual and textual modalities for SL-LLaMA. Second, the SFT stage focuses on training the model to follow the input instruction rather than GPT-4o like Figure 1. By leveraging annotated training data and introducing domain-specific prompts, it enables the model to generate high-quality, contextually, and instruction-compliant response. During this phase, we utilized the following predefined templates as prompts:

"[INST] <sign> *<frames features>* </sign> /n Translate the sign language into German text. [/INST]"

Here, <frame features> represents the encoded visual sequence, while the textual instruction specifies the target task and language. These prompt tokens are passed through the LLM's tokenizer and concatenated with visual tokens to serve as input.

We utilize the efficient fine-tuning method, LoRA [11], for fine-tuning the LLM during SFT. Specifically, for each attention weight matrix W, LoRA defines:

$$W' = W + AB, \tag{4}$$

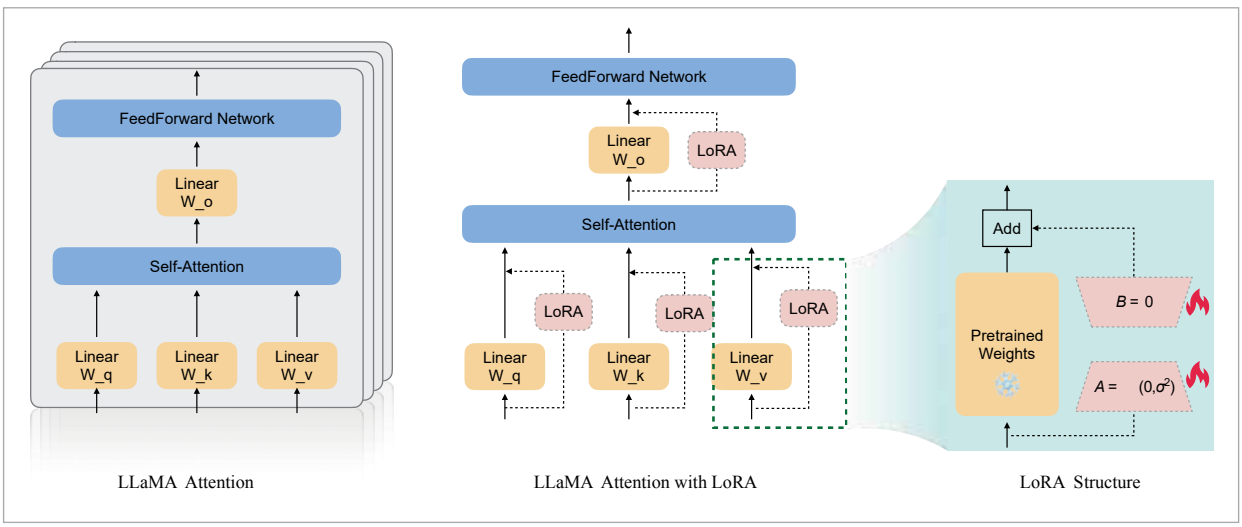
where $A \in R^{d \times r}$, $B \in R^{r \times d}$, and r is the rank parameter controlling adaptation complexity. LoRA introduces low-rank matrices, reducing the number of trainable parameters and computational resources needed [20], making it feasible to train SL-LLaMA in a laboratory setting, as shown in Figure 4. Additionally, LoRA does not alter the original parameters, it preserves the model's general knowledge and performance [26], preventing catastrophic forgetting and overfitting to the limited sign language training data. By the supervised fine-tuning, it ensures robust cross-modal reasoning and fluent textual output, even in resource-constrained scenarios.

3.3. Evaluation Metrics

To compare our method with other state-of-the-art approaches, we utilize the widely used ROUGE and BLEU metrics to evaluate the alignment between the translated content and the reference sentences. ROUGE is a recall-oriented evaluation metric that focuses on assessing the coverage of the generated

Figure 4

Fine-tune the LLM backbone with Low-Rank Adaptor. The pretrained weights of the LLM remain frozen, while low-rank matrixes (denoted as A and B) are introduced to adaptively learn task-specific updates. This approach efficiently injects new knowledge with minimal additional parameters, preserving the LLM's original capacity and reducing the risk of overfitting.



text. BLEU measures the precision of the generated text by calculating the exact match count of N-grams between the generated text and the reference text.

4. Experiments and Discussion

4.1. Dataset and Implementation

The PHOENIX-2014T [3] dataset is a commonly used benchmark dataset for German sign language translation, primarily consisting of televised German weather forecasts. It features 9 sign language performers, and the dataset includes 1,066 sign language vocabulary words and 2,887 translated German spoken language words. The dataset is divided into training, validation, and test sets, comprising 7,096, 519, and 642 sign language videos, respectively. CSL-Daily [41] is a specialized dataset designed in a controlled laboratory environment, encompassing a broad spectrum of daily activities such as shopping, travel and family life.

Due to computational resource constraints, we conduct experiments using two LLMs, llama 2-7B and 13B, and utilize the pre-trained LLaMA tokenizer to encode the text. The minimum rank for LoRA is set to 8. During the alignment learning stage, the batch size is set to 4, and the training lasts for 3 epochs. In the SFT stage, the batch size is set to 2, and the training lasts for 2 epochs. This setup leads to our experimental conclusion that the LLM does not require numerous training cycles; otherwise, performance does not improve and there is a risk of overfitting. All experiments, including training and validation, are conducted on two A100 GPUs with 80GB of memory each.

4.2. Main Results and Analysis

4.2.1. Results on Phoenix-2014T

In Table 1, we conduct a comparative analysis of SL-LLaMA against the current state-of-the-art models on PHOENIX-2014T. The compared models for sign language translation all operate in a gloss-free manner. When the LLaMA-7B model utilized in SL-LLaMA, compared to other methods that do not use a pre-trained language model, our approach demonstrates significantly superior performance across evaluation metrics. In contrast, Sign2GPT, which leverages a small pre-trained language model, achieves competitive results. While our 7B parameter model approaches but does not surpass Sign2GPT, it employs a specialized training strategy for modeling sign language visual feature. When we further increase the LLM parameter size from 7B to 13B, our SL-LLaMA outperforms all methods in the comparison table. These experimental results demonstrate the effectiveness, advancement, and scalability of the proposed SLT framework.

Table 1
Comparison of gloss-free results on PHOENIX-2014T Test Set.

Method	ROUGE	BLEU1	BLEU2	BLEU3	BLEU4
CSGCR (2021) [39]	38.85	36.71	25.4	18.86	15.18
SLTR-T (2020) [5]	-	45.34	32.31	24.83	20.17
GASLT (2023) [36]	39.86	39.07	26.74	21.86	15.74
GFSLT (2023) [40]	40.93	41.39	31	24.2	19.66
VL-Mapper (2023) [6]	46.67	-	-	-	21.36
SignBERT+ (2023) [13]	44.89	44.35	32.09	24.92	20.41
SignFormer (2024) [35]	46.24	-	-	-	20.02
Sign2GPT (2024) [32]	45.23	45.43	32.03	24.23	19.42
Sign2GPT+PGP (2024) [32]	47.11	47.06	33.61	25.85	20.93
SL-LLaMA 7B (Ours)	45.87	45.7	30.68	24.59	20.86
SL-LLaMA 13B (Ours)	49.34	48.95	33.63	26.07	23.73

4.2.2. Results on CSL-Daily

In Table 2, we present the result on the CSL-Daily dataset. The results exhibited a similar trend. The SL-LLaMA 7B model scored 1.19 points lower than sign2GPT on the BLEU4 metric, but it demonstrated a significant advantage over other methods that did not utilize pre-trained language model. The 13B model, due to its increased parameter count and expressive capacity, outperformed both the SL-LLaMA 7B and sign2GPT methods.

Table 2

Comparison of gloss-free results on CSL-Daily Test Set.

Methods	ROUGE	BLEU4
GASLT (2023) [36]	20.35	4.07
NSLT (2018) [3]	34.54	7.56
GFSLT (2023) [40]	35.16	9.88
Sign2GPT (2024) [32]	41.12	12.96
Sign2GPT+PGP (2024) [32]	33.39	9.73
SL-LLaMA 7B (Ours)	38.29	11.77
SL-LLaMA 13B (Ours)	41.96	13.42

4.2.3. Results Analysis

Evaluation results on two benchmarks demonstrate that our proposed SLT framework achieves the best current level in Gloss-free sign language translation. The challenging performance suggests that LLMs can indeed learn advanced grammatical knowledge from visual gestural actions, highlighting their immense potential. In the research on large language models, the scaling law is a critical phenomenon indicating that larger model parameter sizes lead to stronger performance. We observe a similar trend: increasing the parameter of LLM backbone from 7B to 13B results in substantial performance improvement. Based on this observation, it is reasonable to infer that the stronger the fundamental model's capabilities, the greater the performance gains for SLT task. This suggests a promising direction for future work to further enhance translation performance.

Our SL-LLaMA-7B model approaches but does not surpass Sign2GPT, which employs a much smaller 2B language model. Sign2GPT utilizes a pseudo-glosses pre-training strategy to specifically enhance the sign language visual encoder for SLT. This novel approach enables the model to capture visual-semantic information with greater precision, resulting in strong performance. This insight highlights the importance of accurate and comprehensive extraction of visual features for achieving more precise sign language translation. However, it is worth noting that such task-specific optimizations are less transferable to other video-text tasks. In contrast, our framework and training strategy are more flexible and easily adaptable to a broader range of multimodal tasks, providing greater versatility and applicability beyond SLT.

4.3. Ablation Study

Ablation studies primarily analyze the effectiveness and impact of the key modules of SL-LLaMA, the MM-Adaptor, the LLM, and the Vision Encoder on sign language translation.

4.3.1. The Effectiveness of MM-Adaptor

Initially, we build a single Linear layer as the MM-Adaptor, similar to Mini-GPT4. However, we observe that the evaluation metrics significantly lag behind current SoTA methods. Therefore, we employ a two-layer MLP with a nonlinear activation function, GELU, which substantially improves model performance. To determine whether the gains are due to the increased number of parameters or the nonlinear function, we train another two-layer MLP without an activation function. As shown in Table 3, the experimental results indicate that the nonlinear activation function is the primary contributing factor, leading to an improvement of 2.3 to 3.3 points. Our analysis suggests that nonlinear activation functions effectively map visual signals to the textual feature space, as stated in [17]. The research of the multimodal large language model LLaVA also demonstrates the necessity of nonlinear activation functions in cross-modal alignment [19]. In contrast, Linear layers can only perform linear mappings, which makes cross-modal transformations challenging. The success of Mini-GPT4 lies in its use of a Q-Former to process visual features before the Linear layer, incorporating prior knowledge of visual-text cross-modality. Conversely, without the activation function, the impact of the number of Linear layers (parameter number) is negligible, and it even shows a decline in the 13B model.

Table 3
The effective of MM-Adaptor on PHOENIX-2014T test set.

Model	MM-Adaptor	Activation Function	ROUGE	BLEU4
SL-LLaMA-7B	Linear	None	40.63	18.48
	MLP	None	40.07	18.06
	MLP	GELU	45.87	20.86
SL-LLaMA-13B	Linear	Linear	41.59	20.42
	MLP	MLP	40.15	19.74
	MLP	MLP	49.34	23.73

4.3.2. The Impact of Fine-tuning or Freezing LLM

In this study, we further explore whether the model can effectively perform sign language translation with the frozen LLM, thereby investigating the cross-modal potential of text-pretrained language models. In Table 4, although there is a significant performance drop after freezing LLM, it still outperforms models such as GASLT listed in Table 1. This suggests that the text-pretrained LLMs have the inherent potential to understand and process multimodal information. Further analysis indicates that this potential arises from the rich linguistic knowledge and feature representation capabilities developed during large-scale text pretraining. Harnessing this potential, future research may develop new paradigms for vision-language alignment, and even guide and enhance visual tasks through language models.

4.3.3. The Impact of Vision Encoder on LLM

In this framework, the visual features of sign language are extracted through a visual network. *Does the visual feature of sign language videos significantly impact the effectiveness of sign language translation?* Inspired by SL-Transformer [5] and Ref [15], we employ a pre-trained Inception network [28] as the visual encoder. In the ablation experiment, we replace the visual encoder with models pre-trained only on image datasets like ImageNet. We select classical models such as ResNet, EfficientNet (CNN-based), ViT-base, and Swin-Transformer (attention-based) as visual encoders. Each model was configured with an MM-Adaptor to ensure that the final input feature dimension to the LLM is 4096. It is important to note that the entire model training process followed the procedure depicted in Figure 3, and the visual backbone network is frozen during the training.

The experimental results are shown in Table 5 and Figure 5, indicating that Transformer-based visual

Figure 5
The ROUGE and BLEU4 scores of SL-LLaMA with different visual backbone.

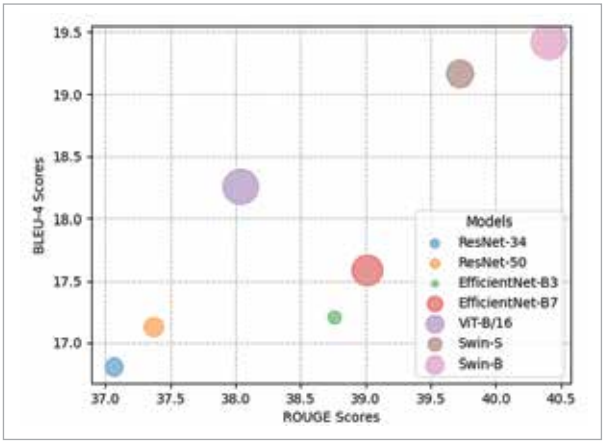


Table 4
The effective of the trainable parameters on PHOENIX-2014T test set. "P" denotes that the parameters are trainable. "O" denotes that the parameters are frozen.

Model	MM-Adaptor	Word Embedding	LLM	ROUGE	BLEU4
SL-LLaMA-7B	P	O	O	36.13	15.95
	P	P	O	37.82	16.12
	P	P	P	45.87	20.86
SL-LLaMA-13B	P	O	O	37.76	16.52
	P	P	O	38.93	16.42
	P	P	P	49.34	23.73

Table 5
The performance of various visual backbones in SL-LLaMA-7B on PHOENIX-2024T test set.

Visual Backbone	Type	#Params	Hidden dims	ROUGE	BLEU4
ResNet-32	CNN	21.8M	512	37.06	16.81
ResNet-50	CNN	25.56M	2048	37.31	17.13
EfficientNet-B3	CNN	12M	1536	38.76	17.21
EfficientNet-B7	CNN	66M	2560	39.01	17.59
ViT-B/16	Self-Attention	86M	768	38.04	18.36
Swin-S	Self-Attention	50M	768	39.72	19.17
Swin-B	Self-Attention	88M	1024	40.41	19.42

Table 6
The results of utilization of local Transformer network on SLT.

Model	Loc. Trans. Net.	Phoenix-2014T		CSL-Daily	
		ROUGE	BLEU4	ROUGE	BLEU4
SL-LLaMA-7B	w/o	41.16	19.27	38.86	11.42
	w	45.87	20.86	38.29	11.77
SL-LLaMA-13B	w/o	47.66	21.43	41.39	13.34
	w	49.34	23.73	41.96	13.42

networks outperformed CNN structures, particularly the Swin Transformer series. Aside from more extensive pre-training, one possible reason is that self-attention architectures have advantages in handling global features, demonstrating a stronger ability to understand high-level semantics, which is suitable for complex tasks. The result of this ablation experiment is consistent with the findings of Ref [5], which used a small-scale model with a classical encoder-decoder Transformer for sign language translation. Therefore, this experiment also suggests that fine-tuning the whole visual encoder during the pre-training phase appears to be necessary when training more general multimodal large models.

4.3.4. The Effectiveness of Local Transformer Network

To evaluate the contribution of the Local Transformer Network (LTN) in modeling temporal dependencies within sign language videos, we conduct the experiment by remove this component from the SL-LLaMA. As Table 6 shown, the results demonstrate the impact of the LTN across two benchmarks. In all the datasets, removing the LTN leads to varying degrees of decrease. This highlights the importance of temporal sequence modeling in capturing nuanced grammatical and semantic transitions inherent in sign language. This perspective is also supported by Sign2GPT, which emphasizes the spatiotemporal representation of sign language, and employs a pseudo-glosses strategy to enhance the vision encoder, achieving outstanding results. Specifically, the BLEU4 scores for SL-LLaMA-7B and 13B decrease by 1.59 and 2.3 on Phoenix-2014T, respectively. Conversely, the impact on the CSL-Daily dataset is less pronounced, with BLEU4 declining by only 0.35 and 0.08. We believe this result is due to the relatively lower baseline scores on the Phoenix dataset, which diminishes the noticeable impact of the LTN. However, the observed trend still demonstrates that the LTN is effective and essential to the overall framework.

Table 7
The results of using various pre-trained LLM as backbone.

Pretrained LLM	#Param	Phoenix-2014T		CSL-Daily	
		ROUGE	BLEU4	ROUGE	BLEU4
Vicuna	7B	43.54	21.17	37.72	10.03
Mistral	7B	44.21	20.95	36.85	11.14
Qwen2	7B	42.87	20.64	40.17	12.96
DeepSeek	7B	43.15	19.34	39.26	12.32
Chinese-LLaMA	7B	43.28	20.03	41.17	13.16

4.3.5. Generalization of Different LLM Backbones

The experiment demonstrates the strong capabilities in gloss-free SLT tasks of our proposed model framework. While the LLaMA series has shown promising results in this context, exploring the effects of using different LLMs as the backbone is crucial to validate the framework's generalization ability, scaling, and robustness. This section evaluates the framework with various LLMs, analyzing the impact on translation performance and the suitability.

To evaluate the generalization of proposed framework, we replace the original LLaMA with five alternative pre-trained LLMs: Vicuna[1], Mistral[2], Qwen2-7B-Instruct[3], DeepSeek[4], and Chinese-LLaMA[5], each configured with 7 billion parameters. These models are selected for their diverse pre-training datasets and linguistic capabilities, and all of them are the mainstream LLMs. Other components of the framework, including the visual encoder and MM-Adaptor, remain unchanged to ensure experimental consistency.

Table 7 presents the experimental results for each LLM backbone. Overall, various LLMs achieve competitive results, validating the framework's adaptability. The proposed SLT framework maintains stable performance across all LLMs, demonstrating its ability to generalize without substantial modification. This highlights its potential for broader multimodal applications. More specifically, on the Phoenix benchmark, Vicuna achieves a higher BLEU4 score, while Mistral obtains better ROUGE. This indicates that in SLT task, Vicuna-based models are more adapted at phrase-level alignment, whereas Mistral-based model excels in capturing global semantic consistency. In addition, the BLEU4 score of Vicuna-base model surpasses that of SL-LLaMA, likely due to Vicuna's extended pretraining beyond the LLaMA, resulting in a stronger foundational model.

On the other hand, we observe language dependency in SLT task. Due to difference in training corpus and linguistic biases, the Chinese proficiency of Vicuna and Mistral is relatively weaker, leading to noticeable performance gaps compared to other LLMs on Chinese sign language translation benchmark, CSL-Daily. This observation suggests that the fundamental model's capabilities can have varying impacts on specific tasks, depending on the linguistic and contextual requirements.

4.3.6. Sensitivity Analysis of Hyperparameters

To validate the robustness and empirical effectiveness of our hyperparameter choices, we conduct a sensitivity analysis on two critical parameters: the LoRA rank used for efficient adaptation of the LLM, and the window size used in the Local Transformer Network (LTN) for video temporal modeling.

The LoRA rank determines the dimensionality of the low-rank matrices used for injecting task-specific knowledge into the frozen LLM. We test ranks of 4, 8, 16, and 32. As shown in Table 8, when the rank is set as 8, achieving a BLEU4 score of 20.86 and a ROUGE score of 45.87. While higher ranks such as 16 and 32 yield similar or slightly fluctuating results, they do not produce consistent improvements. For instance, BLEU4 slightly decreased at rank=32 compared to rank=16. This suggests that higher-rank matrices

[1] https://huggingface.co/lmsys/vicuna-7b-v1.5
[2] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
[3] https://huggingface.co/Qwen/Qwen2-7B-Instruct
[4] https://huggingface.co/deepseek-ai/deepseek-llm-7b-chat
[5] https://huggingface.co/hfl/chinese-llama-2-7b

Table 8
Effect of different LoRA ranks on PHOENIX-2014T test set.

Model	LoRA rank	ROUGE	BLEU4
SL-LLaMA-7B	4	44.39	18.48
	8	45.87	20.86
	16	45.36	20.98
	32	45.79	20.07

Table 9
Effect of different Window sizes on PHOENIX-2014T test set.

Model	Window Size	ROUGE	BLEU4
SL-LLaMA-7B	4	44.45	19.64
	8	45.87	20.86
	12	43.73	20.02
	16	44.16	19.42

may introduce parameter redundancy without clear performance gains. In contrast, rank=8 provides a good trade-off between performance and training efficiency, consistent with findings in prior LoRA-based instruction tuning studies.

We also tested different window sizes in the self-attention mechanism of the LTN: 4, 8, 12, and 16. As shown in Table 9, window size=8 achieved the best overall results, which aligns with our empirical observation that the average ratio between video frames and text tokens is approximately 8:1. Smaller windows (size=4) limit contextual scope and underutilize temporal semantics, leading to reduced performance. Conversely, larger windows (12 or 16) result in performance degradation, especially in BLEU4, which we attribute to semantic interference from adjacent sign gestures. Overly wide attention may include visually similar but semantically distinct signs, reducing the model's ability to extract discriminative features.

5. Conclusion

This paper proposes a framework for sign language translation based on Large Language Models (LLMs). By employing the Alignment-SFT training strategy and LoRA fine-tuning, the model gains the ability to understand visual grammar and generate high-quality text. The model is relatively versatile and can be easily adapted to other vision-to-text tasks. In this framework, we examined the impact of key model components on multimodal tasks. Experimental results demonstrate that high-performance LLMs inherently possess the capability to understand multimodal information, inspiring future research on multimodal large models and data alignment tasks. Beyond sign language translation, the SL-LLaMA framework has potential for various multimodal applications, including video captioning, gesture-based command understanding, and instruction-driven video reasoning. Its modular design allows adaptation to different tasks requiring vision-language alignment under low-resource or instruction-following settings.

6. Limitations and Future Work

While SL-LLaMA demonstrates strong performance on benchmark sign language translation tasks, several limitations remain. First, the current framework relies primarily on manual visual features and does not incorporate non-manual signals such as facial expressions, gaze, or mouthing, which are essential for sign grammar. Future work will explore multi-stream encoding to enrich the visual representation.

Second, although the model generalizes well to two different sign languages, further investigation is needed to extend SL-LLaMA to broader multilingual scenarios. We plan to leverage instruction-based prompting, domain adaptation, and post-training strategies such as reinforcement learning from human feedback (RLHF) to enhance cross-lingual performance.

Finally, real-world testing with the deaf and hard-of-hearing community has not yet been conducted due to logistical constraints. Collaborations with accessibility experts and end users are planned to validate the system in practical settings and guide human-centered refinements.

Acknowledgement

The research was supported in part by the National Key R&D Program of China under No. 2020-JC-JQ-ZD-079-00.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., McGrew, B. GPT-4 Technical Report. arXiv, 2023. DOI: https://doi.org/10.48550/arXiv.2303.08774.
2. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Zhou, J. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv, 2023. DOI: https://doi.org/10.48550/arXiv.2308.12966.
3. Camgoz, N. C., Hadfield, S., Koller, O., Bowden, R. Neural Sign Language Translation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, 7784-7793. https://doi.org/10.1109/CVPR.2018.00812
4. Camgoz, N. C., Hadfield, S., Koller, O., Bowden, R. SubUNets: End-to-End Hand Shape and Continuous Sign Language Recognition. Proceedings of the IEEE International Conference on Computer Vision, 2017, 3056-3065. DOI: https://doi.org/10.1109/ICCV.2017.332
5. Camgoz, N. C., Koller, O., Hadfield, S., Bowden, R. Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, 10023-10033. DOI: https://doi.org/10.1109/CVPR42600.2020.01004
6. Chen, Y., Wei, F., Sun, X., Wu, Z., Lin, S. A Simple Multi-Modality Transfer Learning Baseline for Sign Language Translation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, 5120-5130. https://doi.org/10.1109/CVPR52688.2022.00506
7. Cheng, K. L., Yang, Z., Chen, Q., Tai, Y. W. Fully Convolutional Networks for Continuous Sign Language Recognition. European Conference on Computer Vision, 2020, 697-714. DOI: https://doi.org/10.1007/978-3-030-58586-0_4
8. Chiang, W. L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Stoica, I. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. arXiv, 2024. DOI: https://doi.org/10.48550/arXiv.2403.04132.
9. Du, Y., Xie, P., Wang, M., Hu, X., Zhao, Z., Liu, J. Full Transformer Network with Masking Future for Word-Level Sign Language Recognition. Neurocomputing, 2022, 500, 115-123. DOI: https://doi.org/10.1016/j.neucom.2022.05.051
10. Guo, Z., Hou, Y., Hou, C., Yin, W. Locality-Aware Transformer for Video-Based Sign Language Translation. IEEE Signal Processing Letters, 2023, 30, 364-368. DOI: https://doi.org/10.1109/LSP.2023.3263808
11. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Sifre, L. Training Compute-Optimal Large Language Models. Proceedings of the 36th International Conference on Neural Information Processing Systems, 2022, 30016-30030.
12. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. International Conference on Learning Representations, 2022.
13. Hu, H., Zhao, W., Zhou, W. SignBERT+: Hand-Model-Aware Self-Supervised Pre-Training for Sign Language Understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45, 11221-11239. https://doi.org/10.1109/TPAMI.2023.3269220
14. Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Kivlichan, I. GPT-4o System Card. arXiv, 2024. DOI: https://doi.org/10.48550/arXiv.2410.21276.
15. Koller, O., Camgoz, N. C., Ney, H., Bowden, R. Weakly Supervised Learning with Multi-Stream CNN-LSTM-HMMs to Discover Sequential Parallelism in Sign Language Videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(9), 2306-2320. DOI: https://doi.org/10.1109/TPAMI.2019.2911077
16. Koller, O., Zargaran, S., Ney, H., Bowden, R. Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition. Proceedings of the British Machine Vision Conference, 2016, 136, 1-12. https://doi.org/10.5244/C.30.136
17. Lee, M. Mathematical Analysis and Performance Evaluation of the GELU Activation

Function in Deep Learning. Journal of Mathematics, 2023, 2023, 4229924.https://doi.org/10.1155/2023/4229924

18. Li, D., Xu, C., Yu, X., Zhang, K., Swift, B., Suominen, H., Li, H. TSPNet: Hierarchical Feature Learning via Temporal Semantic Pyramid for Sign Language Translation. Advances in Neural Information Processing Systems, 2020, 33, 12034-12045.
19. Liu, H., Li, C., Wu, Q., Lee, Y. J. Visual Instruction Tuning. Advances in Neural Information Processing Systems, 2024.
20. Liu, S. Y., Wang, C. Y., Yin, H., Molchanov, P., Wang, Y. C. F., Cheng, K. T., Chen, M. H. DoRA: Weight-Decomposed Low-Rank Adaptation. arXiv, 2024. DOI: https://doi.org/10.48550/arXiv.2402.09353.
21. McKinzie, B., Gan, Z., Fauconnier, J. P., Dodge, S., Zhang, B., Dufter, P., Yang, Y. MM1: Methods, Analysis, and Insights from Multimodal LLM Pre-Training. European Conference on Computer Vision, 2024, 304-323. DOI: https://doi.org/10.1007/978-3-031-73397-0_18
22. Núñez-Marcos, A., Perez-de-Viñaspre, O., Labaka, G. A Survey on Sign Language Machine Translation. Expert Systems with Applications, 2023, 213, 118993. DOI: https://doi.org/10.1016/j.eswa.2022.118993
23. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Bojanowski, P. DINOv2: Learning Robust Visual Features Without Supervision. Transactions on Machine Learning Research, 2024.
24. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Lowe, R. Training Language Models to Follow Instructions with Human Feedback. Advances in Neural Information Processing Systems, 2022, 35, 27730-27744.
25. Pu, J., Zhou, W., Li, H. Dilated Convolutional Network with Iterative Optimization for Continuous Sign Language Recognition. International Joint Conference on Artificial Intelligence, 2018, 3, 885-891.https://doi.org/10.24963/ijcai.2018/123
26. Qiu, X., Hao, T., Shi, S., Tan, X., Xiong, Y. J. Chain-of-LoRA: Enhancing the Instruction Fine-Tuning Performance of Low-Rank Adaptation on Diverse Instruction Sets. IEEE Signal Processing Letters, 2024, 31, 875-879. DOI: https://doi.org/10.1109/LSP.2024.3377590
27. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. Language Models Are Unsupervised Multitask Learners. OpenAI Blog, 2019, 1(8), 9.
28. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 2017, 31(1), 4278-4284. https://doi.org/10.1609/aaai.v31i1.11231
29. Team, G., Anil, R., Borgeaud, S., Alayrac, J. B., Yu, J., Soricut, R., Blanco, L. Gemini: A Family of Highly Capable Multimodal Models. arXiv, 2023. DOI: https://doi.org/10.48550/arXiv.2312.11805.
30. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Scialom, T. LLaMA 2: Open Foundation and Fine-Tuned Chat Models. arXiv, 2023. DOI: https://doi.org/10.48550/arXiv.2307.09288.
31. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Fedus, W. Emergent Abilities of Large Language Models. arXiv, 2022. DOI: https://doi.org/10.48550/arXiv.2206.07682.
32. Wong, R., Camgoz, N. C., Bowden, R. Sign2GPT: Leveraging Large Language Models for Gloss-Free Sign Language Translation. arXiv, 2024. DOI: https://doi.org/10.48550/arXiv.2405.04164
33. Xie, P., Cui, Z., Du, Y., Zhao, M., Cui, J., Wang, B., Hu, X. Multi-Scale Local-Temporal Similarity Fusion for Continuous Sign Language Recognition. Pattern Recognition, 2023, 136, 109233. DOI: https://doi.org/10.1016/j.patcog.2022.109233
34. Xie, P., Zhao, M., Hu, X. PiSLTRc: Position-Informed Sign Language Transformer with Content-Aware Convolution. IEEE Transactions on Multimedia, 2021, 24, 3908-3919. DOI: https://doi.org/10.1109/TMM.2021.3109665
35. Yang, E. Signformer Is All You Need: Towards Edge AI for Sign Language. arXiv, 2024. DOI: https://doi.org/10.48550/arXiv.2411.12901.
36. Yin, A., Zhong, T., Tang, L., Jin, W., Jin, T., Zhao, Z. Gloss Attention for Gloss-Free Sign Language Translation. Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition, 2023, 2551-2562. DOI: https://doi.org/10.1109/CVPR52729.2023.00251

37. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L. Sigmoid Loss for Language Image Pre-Training. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, 11975-11986. DOI: https://doi.org/10.1109/ICCV51070.2023.01100
38. Zhang, B., Müller, M., Sennrich, R. SLTUNET: A Simple Unified Model for Sign Language Translation. arXiv, 2023. DOI: https://doi.org/10.48550/arXiv.2305.01778.
39. Zhao, J., Qi, W., Zhou, W., Duan, N., Zhou, M., Li, H. Conditional Sentence Generation and Cross-Modal Reranking for Sign Language Translation. IEEE Transactions on Multimedia, 2021, 24, 2662-2672. DOI: https://doi.org/10.1109/TMM.2021.3087006
40. Zhou, B., Chen, Z., Clapés, A., Wan, J., Liang, Y., Escalera, S., Zhang, D. Gloss-Free Sign Language Translation: Improving from Visual-Language Pretraining. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, 20871-20881. DOI: https://doi.org/10.1109/ICCV51070.2023.01908
41. Zhou, H., Zhou, W., Qi, W., Pu, J., Li, H. Improving Sign Language Translation with Monolingual Data by Sign Back-Translation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, 1316-1325. DOI: https://doi.org/10.1109/CVPR46437.2021.00137
42. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. arXiv, 2023. DOI: https://doi.org/10.48550/arXiv.2304.10592