Hierarchical Feature Alignment for Gloss-Free Sign Language Translation

Sobhan Asasi
Centre for Vision, Speech and Signal
Processing (CVSSP)
University of Surrey
Guildford, Surrey, United Kingdom
s.asasi@surrey.ac.uk

Mohamed Ilyes Lakhal
Centre for Vision, Speech and Signal
Processing (CVSSP)
University of Surrey
Guildford, Surrey, United Kingdom
m.lakhal@surrey.ac.uk

Richard Bowden
Centre for Vision, Speech and Signal
Processing (CVSSP)
University of Surrey
Guildford, Surrey, United Kingdom
r.bowden@surrey.ac.uk

Abstract

Sign Language Translation (SLT) attempts to convert sign language videos into spoken sentences. However, many existing methods struggle with the disparity between visual and textual representations during end-to-end learning. Gloss-based approaches help to bridge this gap by leveraging structured linguistic information. While, gloss-free methods offer greater flexibility and remove the burden of annotation, they require effective alignment strategies. Recent advances in Large Language Models (LLMs) have enabled gloss-free SLT by generating text-like representations from sign videos. In this work, we introduce a novel hierarchical pre-training strategy inspired by the structure of sign language, incorporating pseudo-glosses and contrastive video-language alignment. Our method hierarchically extracts features at frame, segment, and video levels, aligning them with pseudo-glosses and the spoken sentence to enhance translation quality. Experiments demonstrate that our approach improves BLEU-4 and ROUGE scores while maintaining efficiency.

CCS Concepts

• **Computing methodologies → Computer vision**.

Keywords

Sign Language Translation, Vision-Language Pre-training, Large Language Models

1 Introduction

Sign language uses hand gestures, body movements, and facial expressions for communication [21, 22]. Sign language translation (SLT)

aims to convert these visual elements into spoken sentences, enabling communication with non-signers. However, SLT faces challenges in accurately recognizing the components of sign and bridging the gap between visual and textual representations [8, 29].

SLT approaches typically fall into two categories: gloss-based and gloss-free. Gloss-based methods rely on intermediate representations, where sign videos are first translated into a sequence of glosses, written representations of signs using corresponding words in a spoken language, before generating spoken sentences. In contrast, gloss-free methods directly map videos to text without an explicit gloss representation [3, 25]. While gloss-based approaches benefit from structured linguistic information, gloss-free methods bypass the need for gloss annotations, making them more flexible but often more challenging. The high cost and effort required for data collection and meticulous gloss annotation have led to increased interest in gloss-free SLT. Addressing this challenge is the primary focus of this paper.

The increasing use of Large Language Models (LLMs) has had a significant impact on this area. As a result, the latest gloss-free models utilize the remarkable language prediction capabilities of LLMs by converting sign language videos into text-like representations that LLMs can process effectively [17, 26, 30]. When LLMs are applied directly to visual features, they become mere transformers with pre-trained weights due to the large shift between the textual representations on which they were trained and the new visual inputs. To bridge this gap, recent studies have introduced a pretraining phase before the SLT task, which aims to align visual and textual representations [26, 30].

Such pre-training strategies can take two forms: pseudo-gloss pretraining and contrastive visual-language pretraining. Pseudo-gloss pretraining generates intermediate text-like representations from sign videos to help LLMs process them more effectively. Contrastive visual-language pretraining, on the other hand, aligns visual and textual representations by learning a shared feature space, often using contrastive learning techniques to improve modality bridging.

In this paper, we are inspired by the hierarchical structure of signs, in which nuanced hand and body movements form glosses, which in turn construct sentences. As shown in Fig. 1, our method extracts features on three levels: frame, segment, and video. To bridge the modality gap and disparity between visual and textual representations, we apply two alignment strategies: pseudo-gloss pretraining, which aligns segment features with pseudo-glosses, and video-language alignment, which brings video features closer to the target spoken sentences. During the downstream SLT task, the

pretraining alignments are removed, and the video features are directly fed into a decoder for final processing. Our proposed pipeline is more scalable than related approaches such as GFSLT-VLP [30], as it removes the need for a decoder and avoids the need to update the text encoder weights during pretraining. The main contributions of this paper are as follows:

- Our proposed model effectively bridges the gap between visual and textual representations on two different levels of sign video features.
- Our experiments demonstrate the effectiveness of our method, achieving improved gloss-free SLT performance on a benchmark dataset. Notably, our approach enhances both BLEU-4 and ROUGE scores, while preserving efficiency.

2 Related Works

2.1 Gloss-based Sign Language Translation

A dominant paradigm in Sign Language Translation (SLT) is the gloss-based approach, which first recognizes signs into glosses and then translates them to text [4, 5, 28]. Recent works have pushed this forward. MMTLB [4] leverages large pre-trained language models for more fluent translation. TS-SLT [5] employs a dual-stream visual encoder (RGB and keypoints) for more robust feature extraction, and SLTUNET [28] unifies related sub-tasks into a single model to improve learning efficiency and create more universal representations.

2.2 Gloss-free Sign Language Translation

While gloss-based Sign Language Translation (SLT) methods are highly effective due to intermediate supervision from annotated glosses [5, 28], they are limited by the costly and time-consuming nature of gloss creation, which is not feasible for all languages. To address this limitation, gloss-free SLT learns directly from video-text pairs. This approach is more scalable but contends with a larger modality gap, which typically results in lower performance. To address this gap, CSGCR [29] predicts possible language words, generates sentences based on these predictions, and selects the most appropriate sentence using cosine similarity with sign language images. Recently, GFSLT-VLP [30] introduced a pre-training method that aligns sign language images with spoken sentences, converting them into text-like features familiar to LLMs. Leveraging this advancement, subsequent gloss-free models focus on generating more comprehensible features for LLMs. FLa-LLM [6] introduces a two-step approach: initially training on visual information from sign language images using a lightweight translation model and fine-tuning the LLM for SLT. Sign2GPT [26] pre-trains a sign encoder by aligning visual features with prototypes under the supervision of pseudo-glosses, words from spoken sentences via Parts-Of-Speech (POS) tagging, and then utilizes it for SLT.

2.3 Weakly supervised gloss-free SLT

While all gloss-free methods train directly on video-text pairs, weakly supervised approaches are distinct because they introduce an explicit, intermediate mechanism that uses the target sentence to create a proxy for alignment, rather than relying on a standard architecture to learn this implicitly. For example, TSPNet [12]

achieves this architecturally, using a temporal-semantic pyramid to structure the visual learning process. GASLT [27] creates a soft alignment by using a 'Gloss Attention' mechanism, where the semantic content of the target sentence guides the model to focus on the most salient visual frames. More directly, VAP [10] introduces a dedicated pre-training stage to explicitly align visual features with text embeddings from the parallel sentence, bridging the modality gap before the final translation task. These engineered mechanisms are the key feature that defines this weakly supervised category.

2.4 Vision-Language Pre-training

Visual-language pretraining (VLP) models are categorized into single-stream and dual-stream approaches. Single-stream models use a unified transformer encoder for both modalities, while dual-stream models process visual and text inputs separately, aligning them through contrastive learning or cross-attention. GFSLT-VLP [30] was the first to apply CLIP-based dual-stream pretraining to gloss-free SLT but lacks frame-level supervision. VAP [10] addresses this limitation by learning alignments between video frames and spoken words using weak supervision.

3 Methodology

In the following subsections, we present our model architecture (Fig. 1) and describe the training strategy used for SLT.

3.1 Architecture

Visual Encoder. Our video-based sign language translation framework is built around an image encoder that extracts spatial features F from input video frames $X = \{x_0, x_1, \ldots, x_{T^*}\}$, where T^* represents the total number of frames. These features, with a dimensionality of D^*, are structured as $F \in \mathbb{R}^{T^* \times D^*}$ and form the foundation for segment-level feature construction. We use DinoV2 [18], a self-supervised Vision Transformer, as our image encoder. By fine-tuning the ViT-S/14 variant, we adapt it to sign language translation datasets, improving feature extraction and task performance. The class token's output serves as a feature vector, which is transformed linearly and batch-normalized to obtain $F \in \mathbb{R}^{T^* \times D^*}$, before being passed into the temporal encoder.

Temporal Encoder. We use frame features F as input to our temporal encoder to learn spatio-temporal sign representations. Since sign language sequences often contain hundreds of frames, we design a spatio-temporal transformer inspired by prior SLT approaches, incorporating two key modifications for efficiency and effectiveness. Similar to [26], we integrate local self-attention with a window size of seven, a proven technique for SLT [27]. Combined with downsampling, this extends the model's temporal receptive field while reducing redundancy, ensuring compatibility with the LLM encoder. The resulting segment features are represented as $Z \in \mathbb{R}^{T \times D}$, where T is the downsampled sequence length and D is the new feature dimension. To effectively model both temporal and spatial information, we use Rotary Position Embedding (RoPE) [23] for positional encoding. RoPE encodes positional information into attention, preserving temporal and spatial relationships, enhancing the model's ability to capture spatio-temporal structure in sign language sequences.
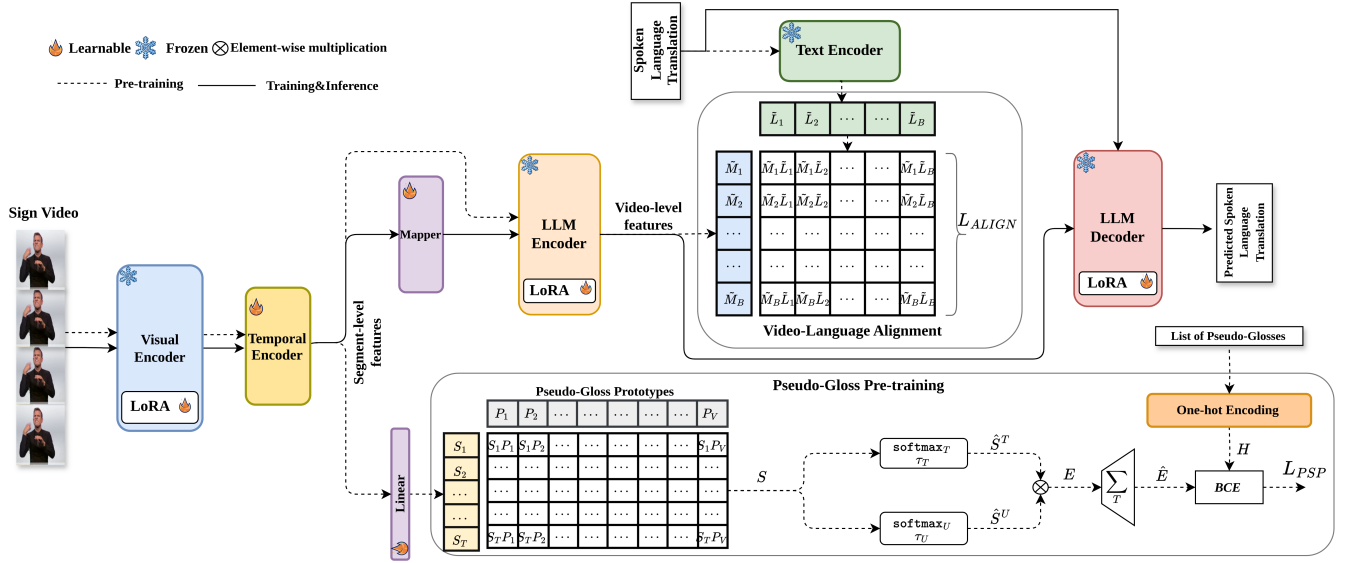
Figure 1: Model overview. Our proposed pipeline we propose is divided into two phases: a comprehensive pre-training phase (dashed line in the figure) and a fine-tuning phase (solid line in the figure). During pre-training, the model bridges the gap between visual and textual representations at two levels. The pseudo-gloss pre-training matches visual features at the segment level with corresponding pseudo-glosses, enabling a fine-grained mapping between video segments and linguistic representations. At the same time, video-to-language alignment uses contrastive learning to map video-translation pairs at a global level, ensuring a coherent alignment between the video and target spoken sentence. During fine-tuning, segment-level features are processed by a mapper to improve their adaptation to the SLT task and boost overall performance.

LLM Encoder. To the refine segment-level features into more expressive representations, we employ a LLM encoder, ψ_{enc}, which transforms S into video-level features $M \in \mathbb{R}^{T \times D}$. Specifically, we use the mBART encoder [14], a 12-layer model initialized with pre-trained parameters from a large corpus. To adapt it to the sign language data while retaining the pre-trained knowledge, we apply LoRA [9], a parameter-efficient fine-tuning method that introduces low-rank matrices into the weights of the model.

3.2 Pretraining and Alignment

Pseudo-Gloss Pretraining. Our goal is to train the image and temporal encoders to learn visual-linguistic representations and ensure that the segment-level features capture the gloss-level information. We extract pseudo-glosses from target outputs using Parts-of-Speech (POS) tagging, keeping key word categories [Noun, Num, Adv, Pron, Propn, Adj, Verb]. The lemmatized German words that meet this filter criterion are treated as pseudo-glosses. Each pseudo-gloss is represented by a 300-dimensional fastText [1] embedding, considered as prototypes, with an additional zero vector for non-sign components, forming the prototype matrix $P \in \mathbb{R}^{D' \times U}$, where $D' = 300$ and U are the embedding size and number of all pseudo-glosses exist in the dataset. As illustrated in Fig. 1, segment-level features $Z = \{z_0, z_1, ..., z_T\}$ are projected to $Z' \in \mathbb{R}^{T \times D'}$ via a linear layer. We compute the cosine similarity between each projected

feature and the pseudo-gloss representations:

$$s_i = \text{sim}(z'_i, P) = \frac{z'_i \cdot P}{\|z'_i\|\|P\|} \quad (1)$$

yielding similarity scores $S \in \mathbb{R}^{T \times U}$, where higher values indicate stronger alignment.

To refine prototype localization, we apply temperature-scaled softmax operations along the temporal and prototype axes:

$$\hat{S}^T = \text{softmax}(S/\tau_T), \quad \hat{S}^U = \text{softmax}(S/\tau_U) \quad (2)$$

where learnable scaling factors τ_T and τ_U control their influence. Element-wise multiplication of $\hat{S}^T$ and $\hat{S}^U$ forms $E \in \mathbb{R}^{T \times U}$, which is aggregated over time to produce $\hat{E} \in \mathbb{R}^U$:

$$\hat{E}_j = \sum_{i=0}^{T} E_{i,j} \quad (3)$$

where high $\hat{E}_j$ values indicate the presence of a pseudo-gloss. We optimize this using binary cross-entropy loss (BCE), assigning 1 if a pseudo-gloss appears in the sign video and 0 otherwise, resulting in an one-hot encoded label $H_j \in \mathbb{R}^U$. As shown in Fig. 1, in the pseudo-gloss pretraining stage, for each input video frames, $L_{\text{psp}} = \text{BCE}(\hat{E}_j, H_j)$.

Video-Language Alignment. Segment-level features S capture visual information aligned with gloss-level representations and are transformed into video-level features via an LLM encoder. To bridge the video-text modality gap, we employ video-language alignment. Inspired by CLIP [20], we use contrastive learning to

Table 1: Experimental results on PHOENIX14T dataset. The best results for gloss-free models are highlighted in bold, while the secondbest results are underlined.

Method	Test Set				
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE
Gloss-based					
MMTLB[4]	53.97	41.75	33.84	28.39	52.65
TS-SLT[5]	54.90	42.43	34.46	28.95	53.48
SLTUNET[28]	52.92	41.76	33.99	28.47	52.11
Weakly supervised gloss-free					
TSPNet[12]	36.10	23.12	16.88	13.41	34.96
GASLT[27]	39.07	26.74	21.86	15.74	39.86
VAP[10]	53.07	-	-	26.16	51.28
Gloss-free					
NSLT[2]	27.10	15.61	10.82	8.35	29.70
NSLT+Luong[16]	29.86	17.52	11.96	9.00	30.70
CSGCR[29]	36.71	25.40	18.86	15.18	38.85
GFSLT-VLP[30]	43.71	33.18	26.11	21.44	42.49
Sign2GPT(w/PGP)[26]	<u>49.54</u>	<u>35.96</u>	<u>28.83</u>	22.52	<u>48.90</u>
FLa-LLM[6]	46.29	35.33	28.03	<u>23.09</u>	45.27
Ours	**49.78**	**36.07**	**29.50**	**23.15**	**49.10**

bring video features closer to the corresponding representations of target spoken sentences and move them away from unrelated representations. Given a target spoken sentence TS_j for video features M_j, we embed it into text features $L_j \in \mathbb{R}^{\overline{T} \times D}$ using a frozen 12-layer mBART encoder ψ_{te} [14], where $\overline{T}$ and D are the number of tokens and dimension of embeddings, respectively. To align video and text, we apply average pooling over frame and token sequences, producing global video $\tilde{M}_j \in \mathbb{R}^D$ and sentence $\tilde{L}_j \in \mathbb{R}^D$ features, as demonstrated in Fig. 1. Contrastive learning is then applied:

$$\mathcal{L}_{\text{align}} = -\frac{1}{2B}\Big(\sum_{j=1}^{B} log \frac{\exp(\text{sim}(\tilde{M}_j, \tilde{L}_j)/\tau)}{\sum_{k=1}^{B}\exp(\text{sim}(\tilde{M}_j, \tilde{L}_k)/\tau)}$$
$$+ \sum_{j=1}^{B} log \frac{\exp(\text{sim}(\tilde{L}_j, \tilde{M}_j)/\tau)}{\sum_{k=1}^{B}\exp(\text{sim}(\tilde{L}_j, \tilde{M}_k)/\tau)}\Big) \quad (4)$$

where $\text{sim}(x, y)$ is the cosine similarity, and τ is a learnable temperature parameter. The final pre-training loss is:

$$\mathcal{L}_{\text{pre-train}} = \mathcal{L}_{\text{align}} + \lambda \mathcal{L}_{\text{psp}} \quad (5)$$

where λ controls the balance between alignment and prototype supervision losses.

3.3 Sign Language Translation

For end-to-end gloss-free SLT, we initialize with pre-trained networks and introduce a linear mapper after the temporal encoder for better adaptation. Given a sign language video V_i, we extract video-level features M_i, which are then processed by a 12-layer mBART decoder [14] ψ_{dec} to generate the spoken sentence $\widehat{TS}_i = (\widehat{TS}_{i,1}, \ldots, \widehat{TS}_{i,T})$.

ψ_{dec} follows an autoregressive approach, starting with <bos> and generating tokens sequentially until <eos>. Training minimizes

cross-entropy loss between predicted tokens $\widehat{TS}_{i,j}$ and ground truth $TS_{i,j}$:

$$\mathcal{L}_{SLT_i} = -\sum_{j=1}^{T} log\, p(\widehat{TS}_{i,j}|TS_{i,1:j-1}, V_i) \quad (6)$$

4 Experiments and Results

4.1 Dataset and Metrics

Dataset. Following previous studies, we evaluate our method with the PHOENIX14T SLT benchmark [2], a dataset in German sign language with three years of news and weather forecasts. It includes 7,096 training, 519 development and 642 test videos.

Evaluation Metrics. We evaluate SLT using BLEU [19] and ROUGE [13] metrics. BLEU-n assesses n-gram precision using the geometric mean, while ROUGE calculates the F1 score based on the longest common subsequence, taking into account word order and position.

4.2 Model and Implementation details

Model. To address memory constraints, we apply LoRA adapters to the top three layers of the visual encoder. The temporal encoder is a four-layer transformer with a hidden dimension of 512, 8 attention heads and an intermediate size of 2048, with temporal downsampling applied after the second layer. LoRA [9] is applied to the LLM encoder and decoder modules, targeting q_{proj} and v_{proj} layers with a rank of 16, a scaling factor (LoRA alpha) of 32 and a dropout rate of 0.1. Training is optimized with the XFormers [11] library with Flash Attention [7] for improved efficiency.

Pre-training. We conduct respective pretraining tasks on the training sets of the PHOENIX14T datasets. The number of total

Table 2: Sensitivity analysis of λ on the SLT results.

Param	BLEU-1	BLEU-2	BLEU-3	BLEU-4
$\lambda = 0.0$	43.71	34.23	27.2	21.82
$\lambda = 0.5$	48.2	35.10	28.43	22.50
$\lambda = 1.0$	**50.28**	**38.07**	**30.48**	**23.68**

Table 3: Number of parameters during pre-training and fine-tuning

Methods	Pre-training		Fine-tuning	
	# Trainable	# Total	# Trainable	# Total
GFSLT-VLP [30]	215,597,633	215,597,633	114,835,008	114,835,008
Sign2GPT [26]	13,889,604	35,946,180	16,689,024	1,771,652,608
Ours	15,200,325	346,123,973	16,819,712	395,924,608

pseudo-glosses is $2,322$ and we initialize the prototype (τ_U) and time temperature (τ_T) to 0.1. Pre-training is conducted for 100 epochs, and the model weights corresponding to the lowest total validation loss ($\mathcal{L}_{\text{pre-train}}$) are saved for fine-tuning.

Fine-tuning. As mentioned before, we introduce a mapper, a single linear layer, to improve the adaptation to the translation task. According to Cambrian-1 [24], a two-stage fine-tuning proves beneficial. In the first stage, all modules except the mapper and the decoder are frozen and training is performed for 100 epochs. In the second stage, we fine-tune all trainable parameters in our proposed method for another 200 epochs. The model is pre-trained and fine-tuned with a batch size of 16 on one A100 GPU. The input sequences are first resized into 256×256, and then randomly/centrally cropped into 224×224 during training/inference. We use the *AdamW* [15] optimizer with a learning rate of 3×10^{-4} and weight decay of 0.001. To enhance regularization, we apply gradient norm clipping with a threshold of 1.0 and use a one-cycle cosine learning rate scheduler with a five-epoch warmup. In addition, data augmentation techniques, including random rotation, resizing, and translation, are applied to all frames within the video sequences with a 50% probability.

4.3 Comparison

Results on PHOENIX14T. Tab. 1 compares gloss-based, weakly supervised gloss-free, and gloss-free methods on the PHOENIX14T test dataset. Our method outperforms other gloss-free SLT models in terms of BLEU-4 and achieves competitive results in all metrics. In particular, it outperforms GFSLT-VLP [30] and Sign2GPT [26] in BLEU-4 score on the test set by +1.71 and +0.63, respectively. Moreover, the improvement in BLEU-4 and ROUGE score highlights the ability of our model to effectively capture long phrases and contextual information. As shown in Tab. 1, our performance still lags behind VAP[10], a weakly supervised method, because it simplifies the translation task by using an intermediate gloss representation. This provides strong supervisory signals that our end-to-end, gloss-free model does not use, forcing it to learn a direct mapping from raw video to text, which is a significantly more complex challenge.

4.4 Ablation Study

We conducted an experiment on the PHOENIX14T development set to evaluate the impact of the parameter λ in Eq. 5 on the final results. As shown in Tab. 2, increasing the weight of the pseudo-gloss pre-training loss ($\mathcal{L}_{\text{psp}}$) within the total pre-training loss ($\mathcal{L}_{\text{pre-train}}$) leads to better performance. We can see in Tab. 3, our model achieves superior efficiency with 15.2 million trainable parameters in pre-training (vs. GFSLT-VLP's 215.6 million) and 395.9 million total parameters in fine-tuning (vs. Sign2GPT's 1.77 billion total parameters), balancing performance, reduced computational load, and adaptability for real-world use.

5 CONCLUSIONS AND FUTURE WORKS

We introduced a hierarchical pretraining strategy for gloss-free SLT, leveraging pseudo-gloss and contrastive video-language alignment to bridge modality gaps. Our method enhances translation performance while maintaining efficiency. Future work can explore further refinements in feature alignment and scalability.

Acknowledgments

References

[1] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics* 5 (2017), 135–146.

[2] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural Sign Language Translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7784–7793. doi:10.1109/CVPR.2018.00812

[3] 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*. 10023–10033.

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

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

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

[7] Tri Dao. 2023. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. arXiv:2307.08691 [cs.LG]

[8] Amanda Cardoso Duarte. 2019. Cross-modal Neural Sign Language Translation. In *Proceedings of the 27th ACM International Conference on Multimedia* (Nice, France) *(MM '19)*. Association for Computing Machinery, New York, NY, USA, 1650–1654. doi:10.1145/3343031.3352587

[9] 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:2106.09685 [cs.CL]

[10] Peiqi Jiao, Yuecong Min, and Xilin Chen. 2025. Visual Alignment Pre-training for Sign Language Translation. In *Computer Vision – ECCV 2024*, Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Springer Nature Switzerland, Cham, 349–367.

[11] Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick

Labatut, Daniel Haziza, Luca Wehrstedt, Jeremy Reizenstein, and Grigory Sizov. 2022. xFormers: A modular and hackable Transformer modelling library.

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

[13] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81.

[14] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. arXiv:2001.08210 [cs.CL]

[15] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101 [cs.LG]

[16] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. arXiv:1508.04025 [cs.CL]

[17] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[18] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. DINOv2: Learning Robust Visual Features without Supervision. arXiv:2304.07193 [cs.CV]

[19] 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 on Association for Computational Linguistics* (Philadelphia, Pennsylvania) *(ACL '02)*. Association for Computational Linguistics, USA, 311–318. doi:10.3115/1073083.1073135

[20] 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*. PMLR, 8748–8763.

[21] Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, Vassilis Athitsos, and Mohammad Sabokrou. 2022. All You Need In Sign Language Production. arXiv:2201.01609 [cs.CV]

[22] Jr. Stokoe, William C. 2005. Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf. *The Journal of Deaf Studies and Deaf Education* 10, 1 (01 2005), 3–37. doi:10.1093/deafed/eni001

[23] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* 568 (2024), 127063.

[24] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. arXiv:2406.16860 [cs.CV]

[25] 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*. 11946–11955.

[26] Ryan Cameron Wong, Necati Cihan Camgöz, and Richard Bowden. [n. d.]. SIGN2GPT: leveraging large language models for gloss-free sign language translation. In *ICLR 2024: The Twelfth International Conference on Learning Representations*.

[27] 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*. 2551–2562.

[28] Biao Zhang, Mathias Müller, and Rico Sennrich. 2023. SLTUNET: A Simple Unified Model for Sign Language Translation. arXiv:2305.01778 [cs.CL]

[29] Jian Zhao, Weizhen Qi, Wengang Zhou, Nan Duan, Ming Zhou, and Houqiang Li. 2022. Conditional Sentence Generation and Cross-Modal Reranking for Sign Language Translation. *IEEE Transactions on Multimedia* 24 (2022), 2662–2672. doi:10.1109/TMM.2021.3087006

[30] 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*. 20871–20881.