Beyond Gloss: A Hand-Centric Framework for Gloss-Free Sign Language Translation

Sobhan Asasi
s.asasi@surrey.ac.uk
Mohamed Ilyas Lakhal
m.lakhal@surrey.ac.uk
Ozge Mercanoglu Sincan
o.mercanoglusincan@surrey.ac.uk
Richard Bowden
r.bowden@surrey.ac.uk

Center for Vision, Speech and Signal Processing (CVSSP)
University of Surrey
Guildford, UK

Abstract

Sign Language Translation (SLT) is a challenging task that requires bridging the modality gap between visual and linguistic information while capturing subtle variations in hand shapes and movements. To address these challenges, we introduce **BeyondGloss**, a novel gloss-free SLT framework that leverages the spatio-temporal reasoning capabilities of Video Large Language Models (VideoLLMs). Since existing VideoLLMs struggle to model long videos in detail, we propose a novel approach to generate fine-grained, temporally-aware textual descriptions of hand motion. A contrastive alignment module aligns these descriptions with video features during pre-training, encouraging the model to focus on hand-centric temporal dynamics and distinguish signs more effectively. To further enrich hand-specific representations, we distill fine-grained features from Hand Mesh Recovery (HaMeR). Additionally, we apply a contrastive loss between sign video representations and target language embeddings to reduce the modality gap in pre-training. **BeyondGloss** achieves state-of-the-art performance on the Phoenix14T and CSL-Daily benchmarks, demonstrating the effectiveness of the proposed framework. Our code is available at https://github.com/elsobhano/BeyondGloss.

1 Introduction

Sign languages use hand and body movement alongside facial expressions, mouth gestures and the complex use of space for communication [41, 43]. Sign language translation (SLT) aims to convert these visual elements into spoken sentences, enabling communication with non-signers. However, due to the inherent complexities of natural signing and the fundamental difference to spoken languages, SLT faces challenges in accurately recognizing the components of sign and bridging the gap between the visual and textual representations [13, 56]. SLT approaches typically fall into two categories: gloss-based and gloss-free. Gloss-based methods [9, 10, 54] rely on intermediate representations, where sign videos are first translated into a sequence of glosses, written representations of signs using corresponding words

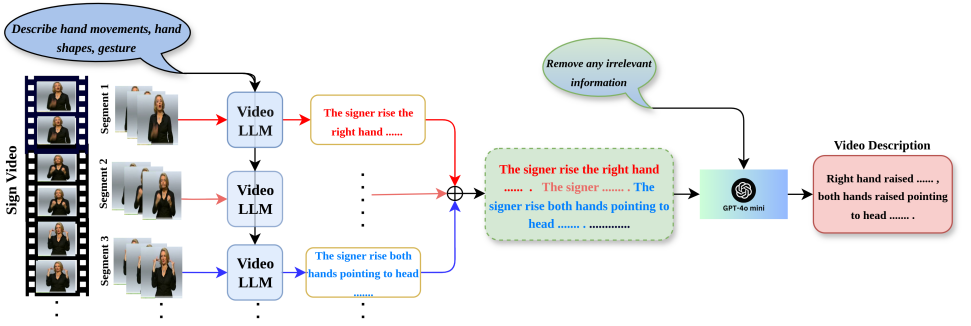

Figure 1: **Proposed Sign Video Descriptor**. The input sign video is first divided into segments that are individually described by a VideoLLM in terms of hand movements, shapes, and trajectories. These segment-level descriptions are then merged and refined by a language model to remove redundancy and preserve temporal order, resulting in a coherent final description that highlights the essential aspects of hand motion across the video.

in a spoken language, before generating spoken sentences. In contrast, gloss-free methods directly map videos to text without an explicit gloss representation [7, 11, 47]. 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 [15, 29, 49]. The increasing use of Large Language Models (LLMs) has had a significant impact on this area [1]. As a result, the latest gloss-free models utilise the remarkable language prediction capabilities of LLMs by converting sign language videos into text-like representations that LLMs can process effectively [11, 49, 57]. 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 pre-training phase before the SLT task, which aims to align the visual and textual representations [3, 49, 57]. Given the central role of hand shape and motion in conveying meaning in sign language, our framework explores the benefit of emphasizing hand-centric features. We achieve this by applying feature distillation from Hand Mesh Recovery (HaMeR) [39], which guides the model to focus more on hand poses and orientations in individual video frames. We then align the video features with a textual description of the entire video, generated using VideoLLMs, which we employ in the context of SLT. These descriptions provide high-level semantic guidance that is more directly aligned with the visual input than conventional spoken language targets. Finally, to enhance performance in the downstream SLT task, we adopt an encoder-decoder architecture based on language models. To further improve the alignment between the encoder's output embeddings and the spoken language targets, we incorporate a contrastive learning strategy inspired by the CLIP framework [40], projecting both video and text features into a shared embedding space.

Our primary contributions can be summarized as: (*i*) Our proposed framework, **Beyond-Gloss**, achieves state-of-the-art results in gloss-free sign language translation on two widely used benchmark datasets. (*ii*) We introduce a novel method for generating hand-centric textual descriptions of sign language videos using VideoLLMs, offering a new form of semantic supervision aligned with visual content. (*iii*) We are the first to leverage HaMeR's transformer-based hand representations in the SLT task. Through feature distillation, we

guide the visual encoder to focus on hand pose and orientation, demonstrating the positive impact of this strategy on translation quality.

The rest of this paper is organised as follows: In Section 2, we review the literature in sign language translation. In Section 3, we outline the proposed **BeyondGloss** approach. Section 4 presents quantitative and qualitative evaluation, whilst Section 5 concludes.

2 Related Works

Gloss-free Sign Language Translation. Gloss-based SLT methods achieve strong performance by using annotated glosses as intermediate supervision [10, 54]. However, glosses are labor-intensive to create and not available for all languages. Gloss-free SLT avoids this by directly translating sign videos into spoken sentences using only video-text pairs. While more scalable, it faces a larger modality gap and typically performs worse than gloss-based methods. To address this gap, CSGCR [56] 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 [57] 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 [11] 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 [49] 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. SignLLM [15] designs a vector-quantization technique to convert sign videos into discrete sign tokens. Lastly, LLaVA-SLT [29] integrates linguistic pretraining, visual contrastive learning, and visual-language tuning within a Large Multimodal Model to effectively bridge the gap between sign language videos and textual translation.

Video Large Language Models. Early VideoLLMs like Video-LLaMA [55] and Video-ChatGPT [35] extended large language models to the video domain by incorporating temporal and multimodal context. Initially limited to short clips and basic temporal understanding, recent models have advanced by combining large-scale video-language data with autoregressive decoding [16, 46]. These systems integrate spatiotemporal encoders with LLMs (*e.g.* LLaMA, GPT) to handle longer sequences and support open-ended video understanding, including tasks like action recognition and captioning. [6, 45]. Although sign language falls under video understanding, it poses unique challenges due to its reliance on fine-grained, frame-by-frame details, especially in hand shapes, movements, and facial expressions. Current VideoLLMs [27, 30, 35, 55] often lose this crucial information due to frame sampling or summarization, and processing all frames at full resolution is computationally costly [48]. To address this, we propose an approach that leverages VideoLLMs for generating detailed hand-centric descriptions suited for sign language understanding.

Vision-Language Pre-training. Existing visual-language pre-training (VLP) models fall into two types: single-stream models, which fuse visual and textual inputs in one encoder using self-attention [24, 25, 28], and dual-stream models, which use separate encoders for each modality and align them via contrastive learning or cross-attention [2, 22, 40]. These pre-trained models have demonstrated remarkable performance in downstream tasks [22, 40]. GFSLT-VLP [57] is the first study to utilize CLIP [40], a pioneering approach in VLP, to tackle the modality gap in gloss-free SLT. Like CLIP, GFSLT-VLP adopts a dual-stream

structure, pre-training the alignment between sign videos and spoken sentences before performing the translation.

Hand Reconstruction Modeling. The typical input to SLT is either RGB video [6, 29] or skeletal based approach [14, 23]. But recent approaches such as HaMeR [39], provide potential alternatives that SLT can be built on. Hand Mesh Recovery (HaMeR) is a fully transformer-based framework for 3D hand reconstruction from monocular images, offering enhanced robustness and accuracy by scaling both model capacity and training data [39]. It utilises a large Vision Transformer and integrates multiple datasets with 2D and 3D hand annotations, achieving state-of-the-art results on standard 3D hand pose benchmarks. Given the central role of hand shapes and movements in sign language communication, HaMeR has recently been adopted for sign language understanding tasks. Notably, Hands-On [18] was the first to apply HaMeR for sign language segmentation. Inspired by this, our approach similarly leverages HaMeR to extract detailed hand features from video frames, aiding in more accurate sign interpretation.

3 Methodology

This section presents our proposed **BeyondGloss** framework. We first describe how high-quality hand features are extracted via distillation in Sec. 3.1, followed by a detailed explanation of our novel Sign Video Descriptor for generating hand motion descriptions in Sec. 3.2. Finally, Sec. 3.3 outlines the overall architecture for SLT.

3.1 Distillation

Distillation [19, 47, 53] is a learning technique where a smaller model (student) is trained to replicate the behavior of a larger, well-trained model (teacher). In feature distillation [20, 42], instead of only imitating final predictions, the student learns to match the intermediate feature representations of the teacher. This helps the student model learn richer internal patterns, improving performance and generalisation with fewer parameters. In our framework, we apply feature distillation to emphasize hand modeling in sign language videos. Specifically, we align the visual encoder's features with those from the HaMeR transformer head to enhance sensitivity to hand shapes and orientations. To ensure that broader spatial information is not lost, we combine the hand-focused features with the original ones, maintaining a balance between detailed hand information and overall scene context.

3.2 Sign Language Video Description

To encourage our framework to focus on hands and their movements, key components of sign language, we aim to align sign video representations with hand-centric descriptions extracted from the videos. However, due to the complexity of hand motions and the inability of off-the-shelf VideoLLMs to capture fine-grained temporal details (as well as memory constraints) [48], we cannot directly use VideoLLMs by simply feeding them full sign videos. To address these challenges, we propose a novel sign video descriptor that leverages the strengths of both VideoLLMs and LLMs. An example of a generated description is provided in Fig 1. As shown in the figure, we begin by segmenting each video into non-overlapping clips of 16 frames. This length is chosen as a balance between temporal resolution and semantic content: segments shorter than 16 frames may lack sufficient motion or variation to be meaningfully described, while longer segments often lead VideoLLMs to produce less

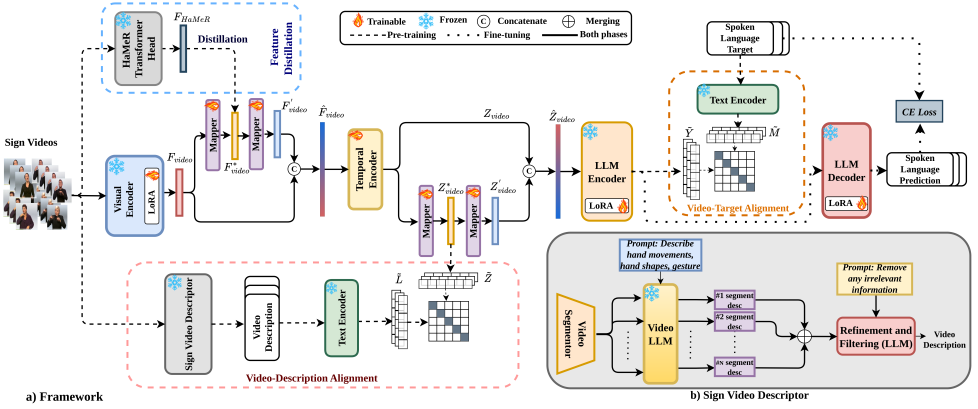

Figure 2: **Overview of BeyondGloss**. a) The framework has two phases: pre-training (dashed+solid lines) and fine-tuning (dotted+solid lines). In pre-training, HaMeR features are distilled, mapped to visual encoder outputs, and fused. After processing by a temporal encoder, the features are mapped to video-description features and fused again. Trainable mappers align feature spaces, removing the need for distillation and alignment modules during fine-tuning. Final features are fed into the LLM encoder and aligned with spoken language, while in fine-tuning, the LLM encoder output is fed directly to the LLM decoder. b) The overview of the proposed Sign Video Descriptor, which segments sign videos, generates descriptions using a VideoLLM, merges them into a sequence, and refines the result through an additional LLM filter.

accurate or overly generalised descriptions. For each segment, we prompt the VideoLLM to describe hand movements, shapes, and trajectories. After obtaining a separate description for each segment, we merge them into a single sequence and pass them to an LLM that refines the content by removing redundant or irrelevant information. This final description preserves the temporal order and captures the most important aspects of hand motion across the entire video. See Appendix for more details and examples.

3.3 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 $V = \{v_0, v_1, \ldots, v_{T^*}\}$, where T^* represents the total number of frames. These features, with a dimensionality of D^*, are structured as $F_{video} \in \mathbb{R}^{T^* \times D^*}$. We use and fine-tune DINOv2 [37] (ViT-S/14 variant), a self-supervised Vision Transformer, as our image encoder. The class token's output serves as a feature vector, which is transformed linearly and batch-normalized to obtain F_{video}.

Feature Distillation. In parallel to the visual encoder, sign videos are input to the HaMeR model to extract hand shape features. Specifically, we utilize the transformer head section of HaMeR, which was originally used to regress hand and camera parameters. In our work, we repurposed the transformer head to extract hand pose and global orientation parameters, as they were highly effective hand shape descriptors for our task. For each frame, the hand pose features are represented as a tensor of $H \in \mathbb{R}^{2\times15\times3\times3}$, while the global orientation is parameterized as $G \in \mathbb{R}^{2\times3\times3}$. For simplicity and computational efficiency, we flatten these parameters, resulting in a 1D feature vector of $H' \in \mathbb{R}^{270}$ for the hand pose and, $G' \in \mathbb{R}^{18}$ for the global orientation. The flattened features are then fused to form $F_{HaMeR} \in \mathbb{R}^{T^* \times 288}$ for each input sign video, where T^* represents the total number of frames. To guide the visual

encoder to extract hand features, we need to map F_{video} to F_{HaMeR} and then combine these mapped features to the original features in order not to lose other important features captured by the visual encoder:

$$F^*_{\text{video}} = \texttt{mapper}_1(F_{\text{video}}) \in \mathbb{R}^{T^* \times 288}, \quad F'_{\text{video}} = \texttt{mapper}_2(F^*_{\text{video}}) \in \mathbb{R}^{T^* \times D}. \tag{1}$$

$$\hat{F}_{\text{video}} = \texttt{concat}(F'_{\text{video}}, F_{\text{video}}) \in \mathbb{R}^{T^* \times 2D}. \tag{2}$$

For hand features distillation to the visual encoder, a simple L_2 loss can suffice. The distillation loss between F^*_{video} and F_{HaMeR} can be represented by:

$$\mathcal{L}_{\text{distill}} = \frac{1}{T^* \times 288} \sum_{i}^{T^*} \sum_{j}^{288} \left\| F^*_{video,i,j} - F_{HaMeR,i,j} \right\|_2. \tag{3}$$

where T^* denotes the number of frames, and 288 is the HaMeR feature dimension. The feature $\hat{F}_{\text{video}}$ is then passed into the temporal encoder for further processing.

Temporal Encoder. We use features $\hat{F}_{video}$ 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. We integrate local self-attention with a window size of seven, a proven technique for SLT [52]. Combined with downsampling, this extends the model's temporal receptive field while reducing redundancy, ensuring compatibility with the LLM encoder. The resulting features are represented as $Z_{video} \in \mathbb{R}^{T \times D}$, 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) [44] for positional encoding.

Video-Description Alignment. To focus on hand-centric temporal dynamics and distinguish signs more effectively, after the temporal encoder, we introduce a contrastive alignment between video and description representations. Rather than relying on a computationally expensive VideoLLM at inference time or discarding temporal information, we project the encoded video features into the description space and then combine them with the original video features to preserve both global and temporal context. As illustrated in Fig. 2, each sign video is processed by a video description module, which generates a textual description. This description is then passed through a text encoder to obtain token-level features, $\hat{L} \in \mathbb{R}^{N \times K}$, where N is the number of tokens and K is the embedding dimension. Simultaneously, the video features Z_{video} are mapped to the same feature space using a two-stage projection:

$$Z^*_{\text{video}} = \texttt{mapper}_3(Z_{\text{video}}) \in \mathbb{R}^{T \times D'}, \quad Z'_{\text{video}} = \texttt{mapper}_4(Z^*_{\text{video}}) \in \mathbb{R}^{T \times D}. \tag{4}$$

$$\hat{Z}_{\text{video}} = \texttt{concat}(Z'_{\text{video}}, Z_{\text{video}}) \in \mathbb{R}^{T \times D''}. \tag{5}$$

The resulting feature $\hat{Z}_{\text{video}}$ are forwarded to subsequent network components. To align video and description features for contrastive learning, we apply a linear projection if $D'' \neq k$, mapping both to a common dimension $\hat{D}$. We then use average pooling across the temporal and token dimensions to obtain global representations $\tilde{Z} \in \mathbb{R}^{\hat{D}}$ for the video and $\tilde{L} \in \mathbb{R}^{\hat{D}}$ for the description. The contrastive loss is computed as:

$$\mathcal{L}_{\text{desc}} = -\frac{1}{2 \times B} \Big(\sum_{j=1}^{B} log \frac{\exp(\texttt{sim}(\tilde{Z}_j, \tilde{L}_j)/\tau_1)}{\sum_{k=1}^{B} \exp(\texttt{sim}(\hat{Z}_j, \tilde{L}_k)/\tau_1)} + \sum_{j=1}^{B} log \frac{\exp(\texttt{sim}(\tilde{L}_j, \tilde{Z}_j)/\tau_1)}{\sum_{k=1}^{B} \exp(\texttt{sim}(\tilde{L}_j, \tilde{Z}_k)/\tau_1)} \Big). \tag{6}$$

where $\texttt{sim}(x,y)$ is the cosine similarity, and τ_1 is a learnable temperature parameter.

Video-Target Alignment. To improve the effectiveness of converting video features into output text, we employ an encoder-decoder-based LLM. The previously constructed feature representation $\hat{Z}_{video}$ is provided as input to the encoder, producing the transformed video features $Y \in \mathbb{R}^{T\times K}$, which are used for subsequent alignment. To address the modality gap between video and language, we adopt a contrastive learning strategy inspired by CLIP [40]. This approach aligns the video features with their corresponding spoken sentence representations, encouraging similarity between matched pairs while distinguishing them from unrelated ones. Given a target sentence TS associated with a video, we encode it into textual features $M \in \mathbb{R}^{\overline{N}\times K}$ using a frozen text encoder, where $\overline{N}$ is the number of tokens and K is the embedding dimension. Average pooling is applied to both $Y_{video} \in \mathbb{R}^{T\times K}$ and M across temporal and token dimensions, respectively, resulting in global video and sentence features: $\tilde{Y} \in \mathbb{R}^{K}$ and $\tilde{M} \in \mathbb{R}^{K}$. We apply a symmetric contrastive loss as follows:

$$\mathcal{L}_{\text{align}} = -\frac{1}{2\times B}\Big(\sum_{j=1}^{B} log\frac{\exp(\texttt{sim}(\tilde{Y}_j,\tilde{M}_j)/\tau_2)}{\sum_{k=1}^{B}\exp(\texttt{sim}(\tilde{Y}_j,\tilde{M}_k)/\tau_2)} + \sum_{j=1}^{B} log\frac{\exp(\texttt{sim}(\tilde{M}_j,\tilde{Y}_j)/\tau_2)}{\sum_{k=1}^{B}\exp(\texttt{sim}(\tilde{M}_j,\tilde{Y}_k)/\tau_2)}\Big). \tag{7}$$

where $\texttt{sim}(x,y)$ is the cosine similarity, and τ_2 is a learnable temperature parameter.
The final pre-training loss is:

$$\mathcal{L}_{\text{pre-train}} = \lambda_{\text{align}}\times\mathcal{L}_{\text{align}} + \lambda_{\text{desc}}\times\mathcal{L}_{\text{desc}} + \lambda_{\text{distill}}\times\mathcal{L}_{\text{distill}} \tag{8}$$

where λ_{align}, λ_{desc} and λ_{distill} control the balance between these three losses. See the Appendix for a comprehensive evaluation conducted to determine the optimal weighting in the pre-training loss.

Sign Language Translation. As illustrated in Fig. 2, the fine-tuning stage excludes the feature distillation, video-description alignment, and video-target alignment components. The mapping modules (purple), introduced in the feature distillation and video-description alignment components, learn to align feature spaces during pre-training. With the mapper modules, the feature distillation and video-description alignment components are no longer required after pre-training, eliminating the need for heavy frozen models such as the HaMeR transformer head, Video-LLM, and LLMs during fine-tuning. This reduction in computational overhead makes our framework more efficient and competitive with other approaches at inference time. During fine-tuning, the only newly introduced module is the LLM decoder, which is attached to the pre-trained LLM encoder and not initialized from pre-training.

Given a sign language video V_i, the decoder aims to generate the corresponding spoken sentence $\widehat{TS}_i = (\widehat{TS}_{i,1},\ldots,\widehat{TS}_{i,T'})$, where T' denotes the number of tokens in the target spoken sentence. The decoder operates in an autoregressive manner; it begins with the special <bos> token and predicts tokens one by one until the <eos> token is reached. The model is trained by minimizing the cross-entropy loss between the predicted tokens $\widehat{TS}_{i,j}$ and the ground truth tokens $TS_{i,j}$ as follows:

$$\mathcal{L}_{SLT_i} = -\sum_{j=1}^{T'} \log\, p(\widehat{TS}_{i,j} \mid TS_{i,1:j-1}, V_i). \tag{9}$$

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets. We conduct experiments on the Phoenix2014T [6] (or Phoenix14T) and CSL-Daily [58] datasets for SLT, evaluating performance on their development and test sets. Phoenix-2014T is a German sign language dataset containing 2,887 unique German words, with 7,096 training samples, 519 validation samples, and 642 test samples. CSL-Daily is a Chinese sign language dataset with a vocabulary of 2,343 Chinese words, comprising 18,401 training samples, 1,077 validation samples, and 1,176 test samples.
Evaluation Metrics. We use BLEU [38] and ROUGE-L [31] to evaluate translation quality. BLEU-n measures n-gram precision [4], and we report BLEU-1 to BLEU-4 scores. ROUGE-L (or ROUGE) evaluates the F1 score based on the longest common subsequence between predicted and reference translations. For simplicity, we denote BLEU-1 to BLEU-4 as B1, B2, B3, and B4, and ROUGE as R.

4.2 Implementation details

Model. To address memory constraints, we apply LoRA [21] adapters to the top three layers of the visual encoder. The adapters are configured with a rank of 4, a dropout rate of 0.1, scaling factor 4.0, and are applied to both self-attention and feed-forward weights. 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. We adopt the `mBART-large-cc25` model [32], which consists of 12 layers in both the encoder and decoder, as the backbone of our LLM-based encoder-decoder architecture. LoRA [21] 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. All mappers in Fig. 2 are trainable and consist of a linear layer followed by `GELU` activation function. Training is optimized with the XFormers [26] library with Flash Attention [12] for improved efficiency. In pre-training, we use `mBART-large-50` [32] as the text encoder for video-description alignment due to its broader multilingual coverage and robustness to noisy, diverse descriptions. For video-target alignment, we choose `mBART-large-cc25` [32] as the text encoder as it offers more stable embeddings for clean, well-structured target sentences in high-resource languages. We use `ShareGPT4Video-8B` [8] as the video LLM in the pre-training stage to generate detailed and semantically rich hand-focused descriptions from sign language video segments, where each segment consists of 16 frames. We also use `GPT-4o-mini` API [36] to generate a single coherent description, maintaining the temporal order of events and removing irrelevant or unnecessary information.
Implementation. The model is pre-trained for 100 epochs, and fine-tuned for 200 epochs, 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* [33] optimizer with a learning rate of 3×10^{-4} and weight decay of 0.001.

4.3 Results

Results on Phoenix14T and CSL-Daily. Tab. 1 compares gloss-based, and gloss-free methods on the Phoenix14T and CSL-Daily test datasets. Our method outperforms other gloss-free SLT approaches in terms of BLEU-4 and ROUGE.
Qualitative Results. Tab. 2 shows translation examples from Phoenix14T. We compare the reference translations with outputs from GFSLT-VLP [57], the only other publicly available

Table 1: Experimental results on Phoenix14T and CSL-Daily datasets. The best results for gloss-free models are highlighted in bold, while the second-best results are underlined. We abbreviate BLEU-1 to BLEU-4 as B1, B2, B3, and B4, and ROUGE as R.

Method	Phoenix14T					CSL-Daily				
	B-1	B-2	B-3	B-4	R	B-1	B-2	B-3	B-4	R
Gloss-based										
MMTLB [8]	53.97	41.75	33.84	28.39	52.65	53.31	40.41	30.87	23.92	53.25
SLTUNET [36]	52.92	41.76	33.99	28.47	52.11	54.98	41.44	31.84	25.01	54.08
TS-SLT [1]	54.90	42.43	34.46	28.95	53.48	55.44	42.59	32.87	25.79	55.72
Gloss-free										
NSLT+Luong [36]	29.86	17.52	11.96	9.00	30.70	34.16	19.57	11.84	7.56	34.54
CSGCR [34]	36.71	25.40	18.86	15.18	38.85	–	–	–	–	–
GFSLT-VLP [36]	43.71	33.18	26.11	21.44	42.49	39.37	24.93	16.26	11.00	36.44
Sign2GPT [19]	49.54	35.96	28.83	22.52	48.90	41.75	28.73	20.60	15.40	42.36
SignCL [31]	49.76	36.85	29.97	22.74	49.07	47.47	32.53	22.62	16.16	48.92
FLa-LLM [11]	46.29	35.33	28.03	23.09	45.27	37.13	25.12	18.38	14.20	37.25
SignLLM [13]	45.21	34.78	28.05	23.40	44.49	39.55	28.13	20.07	15.75	39.91
LLaVA-SLT [20]	51.20	37.51	29.39	23.43	50.44	52.15	36.24	26.47	20.42	51.26
BeyondGloss	**52.38**	**38.57**	**30.74**	**25.49**	**52.89**	**53.12**	**38.63**	**27.82**	**21.53**	**53.46**

Table 2: Qualitative results of Phoenix14T. Green means totally same as the reference.

Reference:	am tag wechseln sonne und wolken einander ab teilweise ist es auch langere zeit sonnig . (During the day sun and clouds alternate partly it is sunny for a long time)
GFSLT-VLP:	am tag wechseln sonne und wolken einander ab es bilden sich langere zeit viel sonnenschein. (During the day, sun and clouds alternate There is a lot of sunshine for a long time)
BeyondGloss:	am tag wechseln sonne und wolken einander ab zeigt sich die auch langere zeit sonnig . (During the day, sun and clouds alternate, and the weather remains sunny for longer periods.)
Reference:	am tag nur hier und da einige sonnige momente vor allem an den alpen . (During the day only here and there some sunny moments, especially in the Alps)
GFSLT-VLP:	morgen zeigt sich mal die sonne wenn dann vor allem an den alpen. (Tomorrow the sun will show up , especially in the Alps)
BeyondGloss:	und morgen scheint und da einige die sonnige vor allem an den alpen . (Tomorrow will shine and there will be some sunny weather, especially in the Alps .)
Reference:	am freitag scheint abseits der nebelgebiete haufig die sonne . (On Friday the sun often shines outside of the foggy areas)
GFSLT-VLP:	am freitag sobald der nebel weg ist sonnenschein. (On Friday as soon as the fog is gone there will be sunshine)
BeyondGloss:	am freitag scheint abseits der nebelgebiete haufig die sonne . (On Friday the sun often shines outside of the foggy areas)

gloss-free model, and our method. While GFSLT-VLP often captures only a few correct words or produces unrelated sentences, our method generates more accurate translations. These qualitative examples demonstrate the effectiveness of our approach.

4.4 Ablation studies

Analysis of Pre-Training Components Contributions. We conduct experiments to evaluate the impact of the core components of our proposed method. Tab. 3 reports the performance obtained by progressively incorporating each component. As shown in the first row, omitting pre-training and alignment between sign videos and output texts results in significantly lower downstream performance, highlighting the importance of pre-training for SLT. The second row demonstrates that video-target alignment contributes substantially to performance improvement. Rows three and four further show the positive effect of our proposed video-description alignment. In rows four and five, feature distillation also contributes to modest improvements in performance. Finally, combining all components together (row five) yields the best overall performance, achieving state-of-the-art results.

Evaluating Different Visual Encoders. We conduct an additional ablation study to examine the impact of different visual encoders on final performance. As shown in Tab. 4, using DINOv2 [37], a transformer-based visual encoder, yields better results compared to ResNet-18 [17], a convolutional encoder. The best performance is achieved when applying LoRA to the

Table 3: Ablation study on key elements in our framework.

	Feature Distillation	Video-Description Alignment	Video-Target Alignment	B-4	R
(1)	–	–	–	16.31	36.33
(2)	–	–	✓	21.87	44.30
(3)	–	✓	✓	23.47	47.17
(4)	✓	✓	–	22.47	44.87
(5)	✓	✓	✓	**25.68**	**53.85**

Table 4: Ablation study on visual encoder.

Visual Encoder	Trainable	B-4	R
ResNet-18[11]	✓	18.97	40.17
DINOv2[51]	×	20.43	42.75
DINOv2[51]	3 last layers fine-tuned by LoRA	**25.68**	**53.85**

Table 5: Ablation study on LLM.

Model	Num of Layers/Num of parameters	B-4	R
MBart [51]	6 / ~115M	22.50	45.70
MT5-Base [50]	24 / ~580M	24.18	49.87
MBart [51]	24 / ~600M	**25.68**	**53.85**

last three layers of DINOv2, leading to our state-of-the-art results. This is because DINOv2 has not been trained on sign language data, and fine-tuning the last three layers allows it to adapt to our domain while only slightly modifying its weights.

Evaluating Different LLMs. As shown in Tab. 5, we assess the impact of using various LLMs within our framework. Both MT5-Base [50] and MBART (24-layer) [32], which have a comparable number of parameters, yield strong results. MBART slightly outperforms MT5-Base, likely due to its stronger multilingual prior knowledge. These findings highlight the generalisation ability of our approach. Additionally, we evaluate a 6-layer version of MBART, similar to the configuration used in GFSLT-VLP [57], which achieves moderate performance. It is worth mentioning that all the tests in the ablation studies are conducted on the Phoenix14T validation set.

5 Conclusion

In this work, we introduced **BeyondGloss**, a novel gloss-free Sign Language Translation framework that emphasizes hand-centric modeling and leverages the capabilities of VideoLLMs. By generating fine-grained, temporally-aware textual descriptions of hand motion and aligning them with sign video features through contrastive learning, our approach effectively bridges the modality gap between visual and linguistic representations. Additionally, the integration of detailed hand features distilled from HaMeR further enriches the sign video representations. Extensive experiments on Phoenix14T and CSL-Daily demonstrate that **BeyondGloss** achieves state-of-the-art performance, underscoring the importance of structured hand modeling and multimodal alignment in advancing gloss-free SLT.

6 Acknowledgement

This work was supported by the SNSF project 'SMILE II' (CRSII5 193686), the Innosuisse IICT Flagship (PFFS-21-47), EPSRC grant APP24554 (SignGPT-EP/Z535370/1), and through funding from Google.org via the AI for Global Goals scheme. This work reflects only the author's views, and the funders are not responsible for any use that may be made of the information it contains.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and others. Gpt-4 technical report, 2024.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[3] Sobhan Asasi, Mohamed Ilyes Lakhal, and Richard Bowden. Hierarchical feature alignment for gloss-free sign language translation. In *IVA'25: Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*. Association for Computing Machinery (ACM).

[4] Peter F Brown, Vincent J Della Pietra, Peter V Desouza, Jennifer C Lai, and Robert L Mercer. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–480, 1992.

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[6] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7784–7793, 2018.

[7] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033, 2020.

[8] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2024.

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

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

[11] Zhigang Chen, Benjia Zhou, Jun Li, Jun Wan, Zhen Lei, Ning Jiang, Quan Lu, and Guoqing Zhao. 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)*, pages 7071–7081, 2024.

[12] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

[13] Amanda Cardoso Duarte. Cross-modal neural sign language translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 1650–1654, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368896.

[14] Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Lei Xie, and Sanglu Lu. Skeleton-aware neural sign language translation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4353–4361, 2021.

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

[16] Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] Low Jian He, Harry Walsh, Ozge Mercanoglu Sincan, and Richard Bowden. Hands-on: Segmenting individual signs from continuous sequences. In *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–5, 2025. doi: 10.1109/FG61629.2025.11099255.

[19] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3779–3787, 2019.

[20] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

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

[23] Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. Skeleton aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3413–3423, 2021.

[24] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1780–1790, 2021.

[25] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021.

[26] 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. xformers: A modular and hackable transformer modelling library, 2022.

[27] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *CoRR*, 2023.

[28] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[29] Han Liang, Chengyu Huang, Yuecheng Xu, Cheng Tang, Weicai Ye, Juze Zhang, Xin Chen, Jingyi Yu, and Lan Xu. Llava-slt: Visual language tuning for sign language translation, 2024.

[30] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, 2024.

[31] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

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

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

[34] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, 2015.

[35] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, 2024.

[36] OpenAI. Gpt-4 technical report. Technical report, OpenAI, 2023.

[37] 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. Dinov2: Learning robust visual features without supervision, 2024.

[38] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics.

[39] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2024.

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[41] Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, Vassilis Athitsos, and Mohammad Sabokrou. All you need in sign language production. *arXiv preprint arXiv:2201.01609*, 2022.

[42] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[43] Jr. Stokoe, William C. Sign language structure: An outline of the visual communication systems of the american deaf. *The Journal of Deaf Studies and Deaf Education*, 10(1): 3–37, 01 2005. ISSN 1081-4159.

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

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

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[47] Andreas Voskou, Konstantinos P Panousis, Dimitrios Kosmopoulos, Dimitris N Metaxas, and Sotirios Chatzis. 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, 2021.

[48] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. In *European Conference on Computer Vision*, pages 453–470. Springer, 2024.

[49] Ryan Cameron Wong, Necati Cihan Camgöz, and Richard Bowden. Sign2gpt: leveraging large language models for gloss-free sign language translation. In *ICLR 2024: The Twelfth International Conference on Learning Representations*.

[50] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, 2021.

[51] Jinhui Ye, Xing Wang, Wenxiang Jiao, Junwei Liang, and Hui Xiong. Improving gloss-free sign language translation by reducing representation density. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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

[53] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017.

[54] Biao Zhang, Mathias Müller, and Rico Sennrich. Sltunet: A simple unified model for sign language translation. In *The Eleventh International Conference on Learning Representations*.

[55] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, 2023.

[56] Jian Zhao, Weizhen Qi, Wengang Zhou, Nan Duan, Ming Zhou, and Houqiang Li. Conditional sentence generation and cross-modal reranking for sign language translation. *IEEE Transactions on Multimedia*, 24:2662–2672, 2022.

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

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