SCOPE: Sign Language Contextual Processing with Embedding from LLMs

Yuqi Liu*, Wenqian Zhang*, Sihan Ren, Chengyu Huang, Jingyi Yu, Lan Xu†

ShanghaiTech University
{liuyq2, zhangwq2022, rensh2022, huangchy, yujingyi, xulan1}@shanghaitech.edu.cn

Abstract

Sign languages, used by around 70 million Deaf individuals globally, are visual languages that convey visual and contextual information. Current methods in vision-based sign language recognition (SLR) and translation (SLT) struggle with dialogue scenes due to limited dataset diversity and the neglect of contextually relevant information. To address these challenges, we introduce SCOPE (Sign language COntextual Processing with Embedding from LLMs), a novel context-aware vision-based SLR and SLT framework. For SLR, we utilize dialogue contexts through a multi-modal encoder to enhance gloss-level recognition. For subsequent SLT, we further fine-tune a Large Language Model (LLM) by incorporating prior conversational context. We also contribute a new sign language dataset that contains 72 hours of Chinese sign language videos in contextual dialogues across various scenarios. Experimental results demonstrate that our SCOPE framework achieves state-of-the-art performance on multiple datasets, including Phoenix-2014T, CSL-Daily, and our SCOPE dataset. Moreover, surveys conducted with participants from the Deaf community further validate the robustness and effectiveness of our approach in real-world applications.

Code and Supplementary Materials —
https://github.com/Godheritage/SCOPE

Introduction

Sign language is the vital visual language used by the Deaf and hard of hearing. Hence, vision-based sign language understanding provides a communication bridge between the Deaf and hearing communities. Such a bridge should accurately and conveniently convey complex contextual information during communication between us humans, especially for dialogue scenarios.

Currently, the two main tasks in vision-based sign language processing include Sign Language Recognition (SLR) (Jiao et al. 2023; Wei and Chen 2023; Zheng et al. 2023) and Sign Language Translation (SLT) (Zhao et al. 2024; Chen et al. 2022b; Yin et al. 2023). SLR converts visual signals into intermediate gloss sequences (Stokoe Jr 2005), while SLT translates visual signals or glosses into natural language. Yet, we notice that most existing methods, both SLR and SLT, focus on translating one sentence at a time, largely ignoring the contextual information of dialogue scenes. Indeed, it's mainly due to the severe lack of sign language datasets for dialogue scenes with sufficient contextual information. For example, the widely adopted PHOENIX2014 (Koller, Forster, and Ney 2015) and PHOENIX2014T datasets (Camgoz et al. 2018) focus on weather forecasts. The How2Sign dataset (Duarte et al. 2021) addresses everyday scenarios but only contains isolated sign language sentences. The CSL-Daily dataset (Zhou et al. 2021a) contains daily sentences but they lack preceding or following context and are essentially still independent statements. On the other hand, in the field of Natural Language Processing, recent advances (Ouyang et al. 2022; Touvron et al. 2023; qwe 2024) with large language models (LLMs) have demonstrated that contextual information can significantly improve semantic understanding and linguistic abilities. Some recent methods (Gong et al. 2024; Wong, Camgöz, and Bowden 2024) utilize LLMs for sign language understanding. Yet, they still focus on per-sentence translation and fall short of analyzing the contextual information for dialogue scenarios. In a nutshell, both the dataset and methodology for contextual vision-based sign language processing remain far-reaching.

To this end, we introduce *SCOPE*, a contextual sign language recognition and translation approach tailored for the dialogue scenes, as shown in Fig. 1 for overview. Specifically, we first contribute a context-based dataset of Chinese sign language dialogues. Our dataset covers a wide range of both daily and professional conversations like shopping and medical treatment. It includes 59,231 dialogue sequences totaling 72.4 hours. For each sequence, we provide video footage, gloss annotations, and dialogue texts, all by professional Deaf individuals from diverse backgrounds.

Secondly, we provide a strong baseline for vision-based contextual sign language processing, which organically utilizes recent LLMs to extract the contextual information from our unique dataset. For the SLR task, we extract sign motion features from the video footage and then introduce a novel embedding alignment module to align them to the context embeddings from a frozen LLM. Then, we feed these

*The authors contributed equally
†Corresponding author

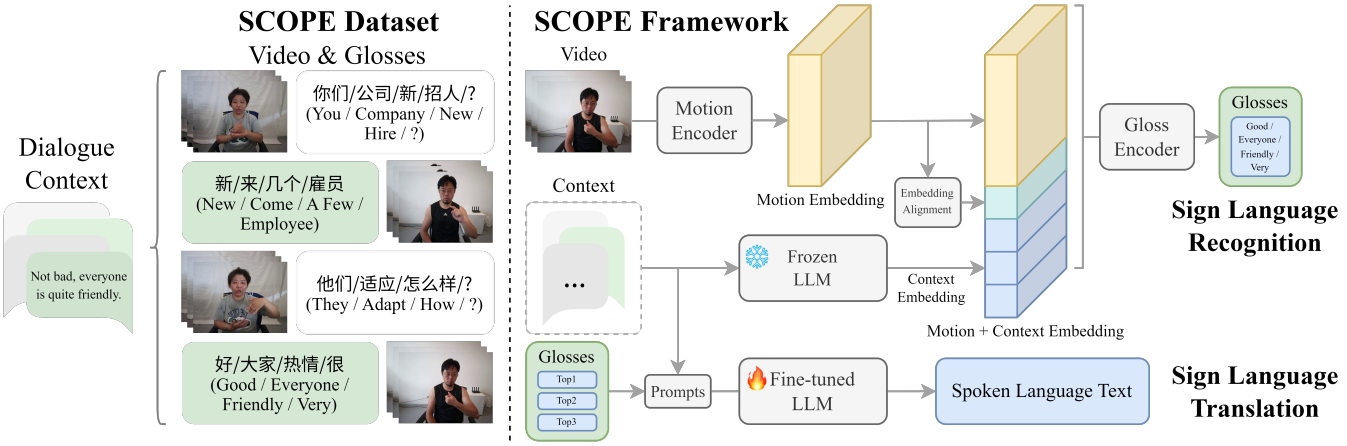

Figure 1: (a) Our SCOPE dataset contains rich contextual information and sign language videos. (b) Our SCOPE framework is a robust context-aware sign language recognition/translation model capable of recognizing dialogue-based sign language gestures, predicting glosses, and generating spoken sentences with the aid of LLMs.

aligned motion/context embeddings into a gloss encoder to obtain the recognized gloss sequences. We observe that such an alignment between the current motions and the preceding contextual information from the LLM is crucial for performance gain. It preserves both the motion and semantic information of the sign language while enabling the concatenation of the contextual embeddings with the input. For the subsequent SLT task, we further leverage the contextual understanding capabilities of the LLM. We use the gloss output from the previous SLR module and the contextual text as inputs and adopt Q-LoRA (Dettmers et al. 2023) to efficiently fine-tune a pretrained LLM model, achieving accurate and natural translations that are closely aligned with the context.

For validation, we conduct comprehensive experiments on both our unique contextual dataset and previously context-free datasets and showcase a companion live demo for sign language translation, which demonstrates the state-of-the-art performance of our approach. In summary, we provide a novel vision-based, context-driven sign language processing approach that utilizes LLMs to address SLR and SLT tasks in dialogue and communication settings. We also contribute a large-scale contextual dataset of Chinese sign dialogues. We believe that both our dataset and baseline approach are the first of their kind to open up the research direction towards context-aware and vision-based sign language analysis. Both our benchmark dataset and baseline approach will be made publicly available.

Related Works

Sign Language Recognition. (SLR) focuses on recognizing glosses from sign videos. While progress has been made in Isolated SLR (ISLR) (Albanie et al. 2020; Tunga, Nuthalapati, and Wachs 2021; Li et al. 2020c; Hu et al. 2021; Li et al. 2020a), current research is shifting to Continuous SLR (CSLR), which converts continuous sign videos into sentence-level gloss sequences. This task involves two main components: feature extraction and translating these features into gloss sequences.

Visual feature extraction often involves extracting features from RGB images using CNNs (Chen et al. 2022a; Li et al. 2020b; Hu et al. 2023c; Min et al. 2021). These features are then modeled with temporal frameworks like RNNs (Camgoz et al. 2018; Ko et al. 2019), LSTMs (Hu et al. 2023a; Cui, Liu, and Zhang 2019), and Transformers (Camgoz et al. 2020; Voskou et al. 2021; Yin and Read 2020) to capture the connection between visual signals and glosses. Some approaches (Zhou et al. 2021b; Papadimitriou and Potamianos 2020) utilize estimated keypoint sequences to describe motions through spatial coordinates or generate heatmaps (Chen et al. 2022b, 2024). However, many methods require video processing, which can be slow and space-consuming, limiting their practical application.

Decoding the extracted features into gloss sequences needs temporal modeling. Hidden Markov Models (HMMs) (Koller, Zargaran, and Ney 2017; Gao et al. 2004; Koller et al. 2016) and Connectionist Temporal Classification (CTC) (Cheng et al. 2020; Zhou et al. 2021b; Min et al. 2021) are commonly used for this purpose. However, most existing methods focus on frame-wise or sentence-wise information, often neglecting the broader linguistic context, resulting in the loss of important language features.

Sign Language Translation. (SLT) aims to translate sign language directly into natural language, bridging the gap between the Deaf community and hearing individuals. This task is challenging due to the modality gap between visual signal and text, compounded by the scarcity of context sign language datasets. Many approaches (Camgoz et al. 2020; Zhou et al. 2021b,a) use SLR results to aid translation. Joint training of SLR and SLT modules has also been explored to improve performance. Some researchers (Li et al. 2020b; Camgoz et al. 2018; Zhou et al. 2023) seek to eliminate gloss by directly translating sign language videos into text using techniques like Conditional Variational Autoencoders and Transformers. SLT involves projecting visual features into coherent textual representations, necessitating insights from both computer vision and natural language processing.

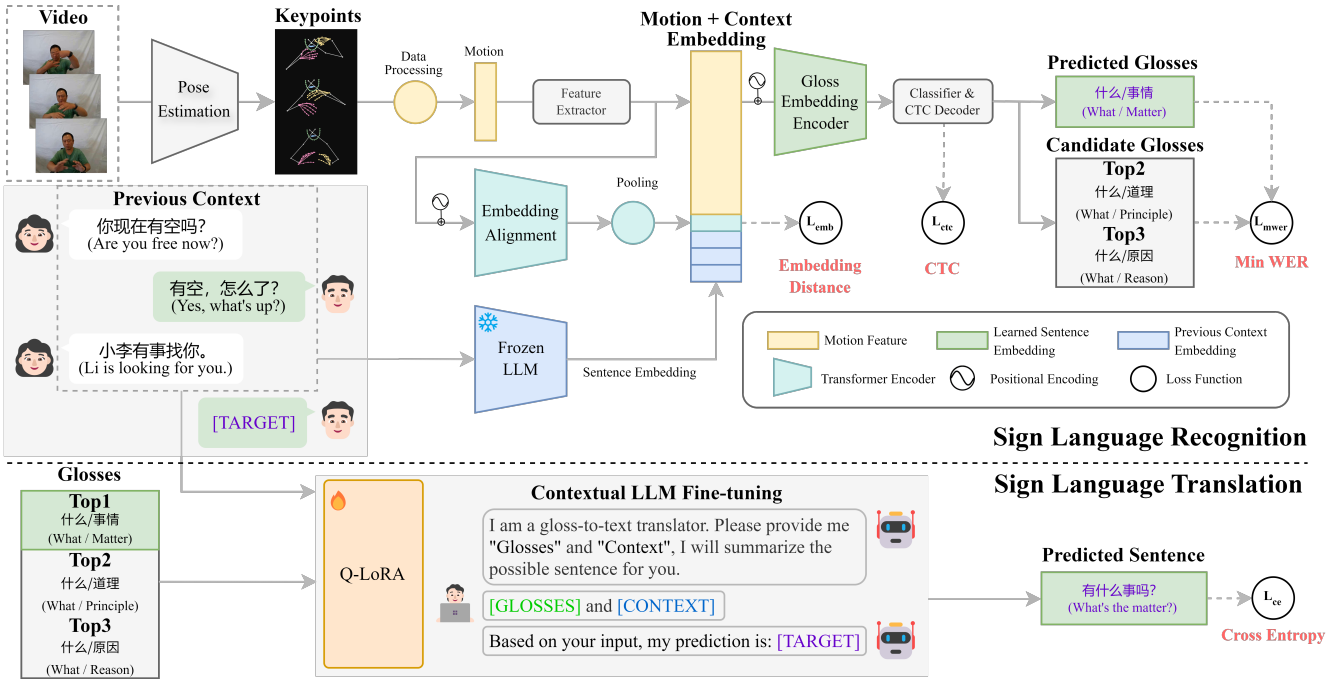

Figure 2: **Overview of SCOPE framework.** Our Embedding Alignment Encoder captures holistic linguistic information from the whole motion sequence. Aligning embedding space to match a frozen LLM enables integrating previous context information for SLR. Finally, Q-LoRA fine-tuning fits an LLM for translating predicted glosses with context into spoken language.

Key advancements leverage pretrained language models like mBART (Liu et al. 2020) for enhanced textual understanding (Chen et al. 2022b,a). Recent studies also explore the use of frozen and fine-tuned large language models (Wong, Camgöz, and Bowden 2024; Gong et al. 2024) to improve translation quality.

Sign Language Dataset. Progress in sign language research has been driven by data. Many researchers have contributed valuable datasets of isolated signs (Wang et al. 2016; Zhang et al. 2016; Joze and Koller 2018; Imashev et al. 2020; Sridhar et al. 2020; Li et al. 2020a; Sincan and Keles 2020; Albanie et al. 2020; Desai et al. 2024). However, while each video clip corresponds to a single sign, the practical utility of such data remains limited.

There are several recent datasets that provide continuous sign language data. For instance, the PHOENIX-2014 (Koller, Forster, and Ney 2015) dataset includes sign language videos from television broadcasts along with corresponding gloss annotations, focusing on weather forecasts. Datasets like SIGNUM (von Agris, Knorr, and Kraiss 2008), PHOENIX-2014T (Camgoz et al. 2018), and CSL-Daily (Zhou et al. 2021a) not only offer gloss annotations but also include natural language translations of the signs, thereby advancing Sign Language Translation (SLT) research. Additionally, the CCSL (Huang et al. 2018) dataset provides images with depth information, increasing the information of sign data, and SeeHear (Albanie et al. 2021) dataset provides a multi-person BSL dataset. The How2Sign (Duarte et al. 2021) dataset stands out with its multi-view information, enabling the capture of 3D sign language motions.

Despite improvements in the size and diversity of sign language datasets, they remain limited in domain coverage. Current corpora consist of context-independent sentences, lacking the contextual relationships needed to fully utilize the linguistic features of sign language, which hinders advancements in SLT research.

Method

We present SCOPE framework, a novel framework that aligns motion features with LLM-provided sentence embeddings of previous contexts, aiming to fully utilize contextually related dialogues in which sign language conversations mainly occur. To address the often overlooked contextual aspects in data collection, we provide SCOPE dataset that annotates sign videos with additional context information, which our model effectively utilizes. Details of SCOPE dataset will be further presented in the Dataset section.

Fig. 2 demonstrates the structure of our SCOPE framework. Our Embedding Alignment Encoder transforms motion features into an embedding that captures the linguistic information of the whole motion sequence. Aligning embedding space to a frozen LLM enables integrating contextual information of previous sentences to recognize glosses. Finally, Q-LoRA fine-tuning fits an LLM for translating predicted glosses into spoken language with the assistance of context information.

Model Details

Embedding Alignment Encoder. We use a transformer encoder structure to extract information from motion features. For the input keypoints $\mathbf{J} = J_1...J_t$, they first pass through the feature extractor linear layer and the temporal sequencer linear layer, which compress the motion information in the spatial and temporal dimensions, respectively, resulting in the intermediate state motion input D, which aligns textual embedding in shape. Next, we need to pretrain an Embedding Alignment Encoder to align features from the motion space with the textual embedding space. The key idea is to directly learn the alignment between the linguistic features of sign language motion and the contextual features of text. In this step, we aim to align the sign motion feature D with the embedding vector of the target sentence. We do this by passing the motion features through the Embedding Alignment transformer encoder and then pooling them to compress the time dimension, resulting in an embedding vector that matches the size of the text embedding. The encoding process, in detail, first embeds the input D into a latent space, represented by h_0, and then obtains the encoded hidden states h_n through N attention layers. Finally, a feed-forward network is used to obtain the encoded vector. The formulas for the transformer motion encoder process are as follows:

$$\begin{aligned} Q = W^Q h_i, K = W^K h_i, V = W^V h_i, \\ h_{i+1} = \textit{Attn}(Q, K, V) = \textit{softmax}(\frac{QK^T}{\sqrt{C}})V, \end{aligned} \tag{1}$$

where W^Q,W^K,W^V are trainable weights, C is the number of channels in the attention layer, and h_{i+1} is hidden states before the next layer.

Supervision by distance to the embedding of the target sentence provided by an LLM.

The loss of the motion encoder is the L2 distance between the pooled embedding vector and the target text embedding vector.

$$\mathcal{L}_{\textit{emb}} = \mathbb{E}||E_{\textit{out}} - E_{\textit{text}}||_2, \tag{2}$$

where L_{emb} denotes the embedding loss, E_{out} is the output of the motion encoder, and E_{text} is the text embedding vector. The text embedding vector is generated by a frozen LLM text embedding model (Neelakantan et al. 2022), which encodes the ground truth sentence meaning of the sign video. Through this process, we align the motion features with the language feature information, enhancing the connections between the isolated sign words.

Gloss Embedding encoder. After aligning the motion features, we obtain an embedding vector that contains both semantic and sign language information. Next, we combine this with the motion features to predict the gloss. For sign language conversations, providing previous language context is crucial for improving the accuracy of recognizing the current target sentence. Therefore, we use a frozen LLM to get the embedding vector for the previous sentences. To minimize irrelevant information, we only keep the last three text embeddings. If there are fewer than three previous sentences, we use a mask to ignore the padding input. The gloss embedding encoder is also a transformer encoder model. The encoding process can be formulated as follows:

$$\begin{aligned} H^0_{t,A} &= \textit{Hidden}(\textit{Cat}(E_t, E_A)), \\ E_{\textit{out}} &= \textit{FFN}(\textit{Attn}(W^Q H^N_{t,A}, W^K H^N_{t,A}, W^V H^N_{t,A})), \end{aligned} \tag{3}$$

where *Hidden* is the hidden layer embedding in the transformer encoder, and *Cat* is the concatenate operation. E_t is the previous stage encoded sequence, and EA is the above text embedding vector. *FFN* is the feed-forward network in the transformer encoder. Passing the output E_{out} through a linear classifier layer, we get the output logic of the glosses.

CTC Decoding. We use connectionist temporal classification (CTC) (Graves et al. 2006) loss to optimize the embedding encoder:

$$\mathcal{L}^y_{\textit{CTC}} = -\textit{logp}(\mathbf{l}|y) = -\textit{log}_{\pi \in \mathcal{B}^{-1}(\mathbf{l})} p(\pi|y), \tag{4}$$

where $\mathbf{l} = l_1...l_t$ is the gloss sequence corresponds to keypoints sequence J. $\mathcal{B}$ is a many-to-one mapping between hypotheses and gloss, and π is the alignment path. In addition, we adopt Minimum World Error Rate (MWER) Training (Meng et al. 2021) technique to reduce the mismatch between training objectives and evaluation metrics, boosting the accuracy of hypotheses on top of the beam. We use beam search during training to decode the top 3 possible gloss sequences. While maintaining the top 1 decoded result as the final output of the SLR network, other candidate glosses contribute to optimization with minimum word error rate (MWER) loss:

$$\mathcal{L}_{\textit{MWER}} = \sum_{n=1}^{N} \bar{P}(\mathbf{Y}^n|\mathbf{J};\theta) R(\mathbf{Y}^n, \mathbf{Y}^*), \tag{5}$$

where $\bar{P}(\mathbf{Y}^n|\mathbf{J};\theta) = \frac{P(\mathbf{Y}^n|\mathbf{J};\theta)}{\sum_{n=1}^{N} P(\mathbf{Y}^n|\mathbf{J};\theta)}$, is the re-normalized posterior over the N-best hypotheses, θ is model parameters, and $R(\mathbf{Y}^n, \mathbf{Y}^*)$ is the number of word errors in a hypothesis $\mathbf{Y}^n$ compared to the reference $\mathbf{Y}^*$.

Furthermore, the top 3 decoded results also serve as the input of the LLM model in SLT task.

Contextual LLM Fine-tuning. Inspired by (Gong et al. 2024), we adopt the idea by using Q-LoRA to fine-tune an LLM as a sign language translator. We adopted the Qwen2 LLM model as our translator. To fine-tune Qwen2, we need to set the LLM using the scenario as a "Sign language translator" and design prompts to guide the model. In the prompts, we provide the top 3 gloss sequences mentioned in and all the above text related to the current sign language sequence, and ask the LLM to summarize the top 3 glosses and guess the correct words to use by checking previous texts. We also provide some summarized task examples to help the LLM understand translation procedures. We use the previous top 3 gloss outputs as input and use the designed prompt along with the above text as auxiliary input, jointly fine-tuning the LLM model. We optimize the model using the cross-entropy loss function:

$$\mathcal{L}_{\textit{llm}} = -\sum_{i} \hat{Y}_t(i) \log(Y_t(i)), i \in N_{\textit{tok}}. \tag{6}$$

$\hat{Y}_t$ is the ground truth textual output, and Y_t is the predicted textual output. N_{tok} is the number of classes in the tokenizer.

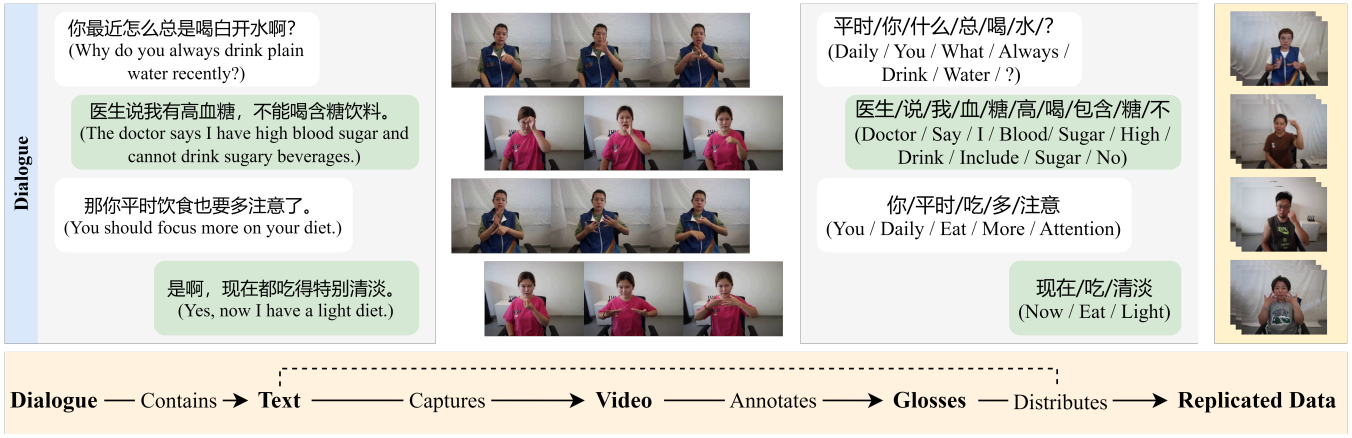

Figure 3: **SCOPE dataset collection pipeline.** Given dialogue texts, experienced signers produce corresponding sign videos along with self-annotated glosses. For each video, other signers replicate data based on the glosses and the text.

Dataset	Language	Videos	Duration(h)	Signers	Vocab	Gloss	Text	Dialogue	Source
PHOENIX-2014	DGS	6,841	8.9	9	1k (German)	✓	×	×	TV
PHOENIX-2014T	DGS	8,257	11	9	3k (German)	✓	✓	×	TV
CSL-Daily	CSL	20,654	22.8	10	2k (Chinese)	✓	✓	×	Lab
How2Sign	ASL	35,191	79	11	16k(English)	✓	✓	×	Lab
Ours	CSL	59,231	72.4	12	5k (Chinese)	✓	✓	✓	Lab

Table 1: **Dataset comparisons.** Key statistics of widely used sign language datasets for comparison. Our dataset is currently the largest dataset in CSL (Chinese Sign Language) that contains dialogue context information.

Data Processing

Iris Normalization. To fetch keypoint sequences, we utilized DWPose (Yang et al. 2023) to obtain whole-body keypoints (COCO-WholeBody (Jin et al. 2020)) from sign language videos. Each keyframe contains 133 keypoints $J = \{J_{1,i}, ..., J_{T,i} | i = 1...133\}$. However, such keypoints are often influenced by the input video resolution and the distance between the person and the camera. A scaling process is needed to mitigate the impact of input distortions on motion. Inspired by (Lugaresi et al. 2019), we choose the length of the lower eyelid in humans as the golden standard, comparing the eyelid length differences to get the scale factor and scale motions to the standard size:

$$J_t^{scaled} = \frac{(J_x, J_y)}{|(J_{x_{63}} - J_{x_{64}})|}, \tag{7}$$

Where J_t^{Scaled} are scaled joints under frame t, (J_x, J_y) are 2D coordinates of joints. $J_{x_{63}} - J_{x_{64}}$ is eyelid length; 63 and 64 are indexes of left and right wings of the eyelid in COCO-WholeBody.

Data Centralizing. After that, we followed (Jiao et al. 2023) by selecting 77 keypoints and dividing them into 5 groups, then applied group-specific centralization to decouple multi-grained motion information in skeleton data:

$$J_{t,k} = J_{t,k} - J_{t,r_g}, k \in G, \tag{8}$$

where $J_{t,k}$ denotes joints under the t frame, group k, G are 5 groups, and r_g is the root keypoint of group g.

Data Standardizing. Finally, we standardize all input motions to make their distribution more closely conform to a standard distribution, which eliminates the difficulties that motion corner cases bring to training:

$$J_i^{std} = \mathcal{N}(J_i - \frac{\sum^{n}\sum^{t} J_i}{N \times T}, \frac{I}{J_i^{std}}), i = 1...77, \tag{9}$$

where J_i denotes the i-th joint, and J^{std} is the standard deviation of joint i.

SCOPE dataset

A sign language dataset with contextual information is essential to fully leverage the power of context in implementing our context-aware approaches. We propose SCOPE, a large-scale Chinese sign language dataset that includes contextual dialogue information. Our data information and dataset comparison can be found in Tab.1.

Data Collection. Our dataset primarily focuses on daily conversations within the Deaf community, as well as dialogues involving specialized terminology in more professional settings (Bragg et al. 2019). Our dataset includes daily subjects such as school experiences, shopping, and social interactions. Glosses encompass specific products and brands, titles of audiovisual works, and other relevant terms. For more details, please refer to the supplementary material.

Data collection is carried out by a team whose primary members are several professional Deaf signers and three

sign language linguistics experts. The team also includes a diverse group of Deaf individuals across various ages, genders, occupations, and educational backgrounds to capture diverse signing styles. To ensure a natural dialogue environment, each sentence was derived from conversations recorded in real situations.

Fig.3 illustrates our data collection pipeline. Professional Deaf signers receive reference sentences and record corresponding sign language videos. Capable of self-annotating recorded motion into glosses, they produce gloss annotations that are distributed to other signers. With sentences and glosses as references, other signers replicate data with diverse signing habits and styles. We ensure that four different signers record videos for each piece of text at a resolution of 640x480 and a frame rate of 30 frames per second.

Annotation Cleaning and Validation. Self-annotated glosses still suffer from inconsistencies across different annotators. A same sign is sometimes annotated with synonyms, while a sequence of signs may get interpreted into phrases or separated words. To mitigate such issues, we follow CSL-Daily (Zhou et al. 2021a) to apply a multi-round data cleaning process with our SCOPE SLR model.

Particularly, we compute Minimum Edit Distance (MED) between predicted glosses and ground truth from annotation. The results enable us to identify patterns of synonyms, word division and word combination. Our sign language linguistics experts then examine frequently confused patterns and correct misannotated data. We iterate such a process to reduce our gloss vocabulary size from over 7k to 5k, significantly improving the dataset's quality.

Experiments

Experimental Setup

Datasets and Evaluation Metrics. For the SLR task, we evaluate our proposed method on PHOENIX14, PHOENIX14-T, CSL-Daily, and SCOPE dataset; the latter three datasets are also utilized in experiments on SLT task. For the number of samples and other details of datasets, please refer to supplementary materials. Train/dev/test splits of the existing datasets are maintained. For our SCOPE dataset, we follow (Zhang et al. 2024) to use widely adopted split ratios to randomly split our dataset by 80%, 5% and 15% into train, dev, and test sets, carefully ensuring that no same sentence appears in different sets and any sentence in the dev set or test set does not appear in context dialogues of the training set.

Following previous works (Chen et al. 2022b), we evaluate SLR the Word Error Rate (WER), which measures the percentage of incorrect words in the recognized text. For SLT, we use BLEU (Papineni et al. 2002) and ROUGE-L (Lin 2004), which assess translation quality based on n-gram overlap and longest common subsequences. Lower WER indicates more accurate recognition results, while higher BLEU and ROUGE-L signify better translations.

Implementation Details. We obtain sentence embeddings by OpenAI's text-embedding-ada-002 (Neelakantan et al. 2022) model. Body 2D keypoints are collected from videos using DWPose (Yang et al. 2023). Our motion feature extractor block consists of a 2-layer MLP with a temporal Conv1D layer. The embedding alignment encoder and gloss encoder are both 8-head transformer encoders with 2 and 4 layers, respectively, with hidden size 1568 and feed-forward size 3136. We adopt the AdamW optimizer and use cosine annealing schedules, with 20 epochs focusing on alignment embedding, and 60 epochs for gloss encoder training while keeping the previous module frozen. When training without the context module, we do not provide context information by filling context embeddings with zeros and providing empty context input for LLM. All experiments are executed on 8 NVIDIA A800 GPUs. More implementation details are provided in the supplementary materials.

Comparison with State-of-the-art Methods

Sign Language Recognition. We evaluate our approach by comparing results on multiple datasets with recent methods SEN-CSLR (Hu et al. 2023c), TwoStream-SLR (Chen et al. 2022b) and CorrNet (Hu et al. 2023b).

On our SCOPE dataset, we evaluate their performance by training their open-sourced framework. We perform preprocessing to match the input specifications of each method and train their models adhering as closely as possible to their proposed training setups.

As shown in Tab.2, our context-free SCOPE outperforms other SLR methods in WER by **2.7**%/**2.2**% on CSL-Daily dev/test sets and **3.3**%/**3.1**% on SCOPE dataset, respectively. Adding context information further improves our model's accuracy by **2.2**%/**3.3**% WER, revealing that contextual understanding effectively assists gloss recognition.

Sign Language Translation. On the SLT task, we compare our approach with state-of-the-art gloss-supervised and gloss-free methods. Similarly, we stick to their respective training configurations when training their models on SCOPE dataset. Results in Tab.3 show that our approach outperforms previous methods by +3.73/+3.57 BLEU and +3.50/+3.14 BLEU in Phoenix-2014T and CSL-Daily dev/test sets. Additionally, our full approach on SCOPE dataset brings another +3.26/+2.79 BLEU improvement, which we attribute mainly to context-aware LLM fine-tuning. Notably, when comparing across datasets, SCOPE dataset generally yields better performance for any fixed method. We primarily attribute this result to our robust data annotation and cleaning process.

Ablation Studies

We conduct ablation experiments for both SLR and SLT tasks to validate the contributions of each component. The comparison between our full and context-free SCOPE model also suffices as an ablation study to demonstrate the significance of context information, both in recognition and in LLM fine-tuning. When the embedding alignment encoder is removed, the context sentence embeddings are concatenated to motion features directly, and $\mathcal{L}_{emb}$ no longer serves as a supervision term. The performance of this model declines by 4.4%WER and 16.01 BLEU, and we note that it takes significantly longer for this model to converge. Thus,

Method	Phoenix-2014		Phoenix-2014T		CSL-Daily		SCOPE	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test
SEN-CSLR (Hu et al. 2023c)	19.5	20.9	19.3	20.7	31.1	30.7	40.2	41.1
TwoStream-SLR (Chen et al. 2022b)	**18.4**	**18.8**	**17.7**	**19.3**	**25.4**	**25.3**	40.8	40.5
CorrNet (Hu et al. 2023b)	18.9	19.7	18.9	20.5	30.6	30.1	33.5	33.8
Ours-SLR* w/o Context	**18.8**	**19.2**	**17.8**	**19.0**	**22.7**	**23.1**	**30.2**	**30.7**
Ours-SLR*	-	-	-	-	-	-	**28.0**	**27.4**

Table 2: **Quantitative evaluation** of **Sign Language Recognition (SLR)** task. WER is adopted as the evaluation metric. We train other methods and ours on our SCOPE dataset. Also, our model without context input is evaluated on other popular datasets. The red and blue entries indicate the best and the second best results.

Dataset	Method	Dev					Test				
		R↑	B1↑	B2↑	B3↑	B4↑	R↑	B1↑	B2↑	B3↑	B4↑
P-2014T	MMTLB-S2T (Chen et al. 2022a)	53.10	53.95	41.12	33.14	27.61	52.65	53.97	41.75	33.84	28.39
	TwoStream-S2T (Chen et al. 2022b)	54.08	54.32	41.99	34.15	28.66	53.48	**54.90**	42.43	34.46	28.95
	CV-SLT (Zhao et al. 2024)	**54.43**	**55.09**	**42.60**	**34.63**	**29.10**	**54.33**	54.88	**42.68**	**34.79**	**29.27**
	Ours* w/o Context	**67.09**	**61.80**	**49.09**	**39.53**	**32.83**	**60.06**	**61.74**	**49.22**	**39.61**	**32.84**
CSL-Daily	MMTLB-S2T (Chen et al. 2022a)	53.38	53.81	40.84	31.29	24.42	53.25	53.31	40.41	30.87	23.92
	TwoStream-S2T (Chen et al. 2022b)	55.10	55.21	42.31	32.71	25.76	55.72	55.44	42.59	32.87	25.59
	CV-SLT (Zhao et al. 2024)	**56.36**	**58.05**	**44.73**	**35.14**	**28.24**	**57.06**	**58.29**	**45.15**	**35.77**	**28.94**
	Ours* w/o Context	**60.18**	**60.37**	**47.21**	**37.36**	**31.74**	**60.68**	**60.48**	**49.61**	**40.01**	**32.08**
SCOPE	MMTLB-S2T (Chen et al. 2022a)	63.25	60.72	50.33	40.39	31.61	64.30	61.69	51.75	41.98	33.56
	TwoStream-S2T (Chen et al. 2022b)	63.40	60.87	50.48	40.74	31.65	64.30	61.78	51.86	42.17	33.50
	CV-SLT (Zhao et al. 2024)	65.71	63.16	52.00	**43.69**	**37.10**	66.06	62.69	52.12	**44.14**	**37.82**
	Ours* w/o Context	**69.34**	**64.31**	**53.15**	43.57	34.83	**69.46**	**64.62**	**53.64**	44.13	35.80
	Ours*	**69.78**	**65.68**	**55.06**	**46.18**	**38.09**	**70.14**	**65.85**	**55.42**	**46.56**	**38.59**

Table 3: **Quantitative evaluation** of **Sign Language Translation (SLT)** task. (R: ROUGE, B: BLEU) We train other methods on our dataset, our method on all three datasets. For non-context-based data, we train our method without context. The red and blue entries indicate the best and the second-best results.

Ablation Study	SLR	SLT	
	WER↓	R↑	B4↑
Full SCOPE*	**27.4**	**70.14**	**38.59**
w/o Context	30.7	69.46	35.80
w/o Embedding Encoder	31.8	51.77	22.58
w/o $\mathcal{L}_{MWER}$	37.6	48.64	15.55
w/o Iris Normalization	35.8	51.51	21.76

Table 4: **Ablation studies** of our contextual design and data processing algorithm.

we deduce that the model encounters difficulties in aligning motion features with LLM context embeddings and ultimately behaves poorly. The removal of $\mathcal{L}_{MWER}$ directly causes more word errors, thus deteriorating the translation results. The distribution of raw keypoints is severely biased without our Iris Normalization process, rendering the model overfit to extreme cases and unsuitable for real-time practical use with different aspect ratios and camera resolutions.

Real-time Application and User Studies

Authentic feedback from the Deaf community is the gold standard for practical use. We have developed a real-time SLT application to assist Deaf individuals in accessing dental care. Details are provided in the supplementary materials. We conducted a survey on their user experience, and questions concerning random SLR or SLT results. We collected 40 responses, rating our application user experience as 4.15 / 5 on average, accuracy of SLR results as 3.96 / 5, and SLT results as 3.98 / 5. These ratings indicate a positive response from the Deaf community, providing strong evidence of our research's effectiveness.

Conclusion

We present the SCOPE dataset, the first dialogue-based Chinese Sign Language dataset featuring both gloss and text annotations. This dataset encompasses 72.4 hours of sign language videos collected from professional Deaf groups, complemented by 59,231 text annotations. Building on this dataset, we introduce the SCOPE framework, a robust pipeline designed to address Sign Language Recognition (SLR) and Sign Language Translation (SLT) tasks with rich contextual information. Our comprehensive evaluations demonstrate the effectiveness of our methods and the significant improvements enabled by our dataset for the sign language community. We believe that SCOPE will catalyze future research in context-based sign language processing.

References

2024. Qwen2 Technical Report.

Albanie, S.; Varol, G.; Momeni, L.; Afouras, T.; Brown, A.; Zhang, C.; Coto, E.; Camgöz, N. C.; Saunders, B.; Dutta, A.; et al. 2021. SeeHear: Signer diarisation and a new dataset. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2280–2284. IEEE.

Albanie, S.; Varol, G.; Momeni, L.; Afouras, T.; Chung, J. S.; Fox, N.; and Zisserman, A. 2020. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, 35–53. Springer.

Bragg, D.; Koller, O.; Bellard, M.; Berke, L.; Boudreault, P.; Braffort, A.; Caselli, N.; Huenerfauth, M.; Kacorri, H.; Verhoef, T.; et al. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, 16–31.

Camgoz, N. C.; Hadfield, S.; Koller, O.; Ney, H.; and Bowden, R. 2018. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7784–7793.

Camgoz, N. C.; Koller, O.; Hadfield, S.; and Bowden, R. 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.

Chen, H.; Wang, J.; Guo, Z.; Li, J.; Zhou, D.; Wu, B.; Guan, C.; Chen, G.; and Heng, P.-A. 2024. SignVTCL: Multi-Modal Continuous Sign Language Recognition Enhanced by Visual-Textual Contrastive Learning. *CoRR*, abs/2401.11847.

Chen, Y.; Wei, F.; Sun, X.; Wu, Z.; and Lin, S. 2022a. A simple multi-modality transfer learning baseline for sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5120–5130.

Chen, Y.; Zuo, R.; Wei, F.; Wu, Y.; Liu, S.; and Mak, B. 2022b. Two-Stream Network for Sign Language Recognition and Translation. *NeurIPS*.

Cheng, K. L.; Yang, Z.; Chen, Q.; and Tai, Y.-W. 2020. Fully convolutional networks for continuous sign language recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, 697–714. Springer.

Cui, R.; Liu, H.; and Zhang, C. 2019. A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia*, 21(7): 1880–1891.

Desai, A.; Berger, L.; Minakov, F.; Milano, N.; Singh, C.; Pumphrey, K.; Ladner, R.; Daumé III, H.; Lu, A. X.; Caselli, N.; et al. 2024. ASL citizen: a community-sourced dataset for advancing isolated sign language recognition. *Advances in Neural Information Processing Systems*, 36.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2023. QLoRA: efficient finetuning of quantized LLMs (2023). *arXiv preprint arXiv:2305.14314*, 52: 3982–3992.

Duarte, A.; Palaskar, S.; Ventura, L.; Ghadiyaram, D.; DeHaan, K.; Metze, F.; Torres, J.; and Giro-i Nieto, X. 2021. How2sign: a large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2735–2744.

Gao, W.; Fang, G.; Zhao, D.; and Chen, Y. 2004. A Chinese sign language recognition system based on SOFM/SRN/HMM. *Pattern Recognition*, 37(12): 2389–2402.

Gong, J.; Foo, L. G.; He, Y.; Rahmani, H.; and Liu, J. 2024. Llms are good sign language translators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18362–18372.

Graves, A.; Fernández, S.; Gomez, F.; and Schmidhuber, J. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, 369–376.

Hu, H.; Zhao, W.; Zhou, W.; Wang, Y.; and Li, H. 2021. SignBERT: Pre-training of hand-model-aware representation for sign language recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11087–11096.

Hu, L.; Gao, L.; Liu, Z.; and Feng, W. 2023a. Continuous sign language recognition with correlation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2529–2539.

Hu, L.; Gao, L.; Liu, Z.; and Feng, W. 2023b. Continuous sign language recognition with correlation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2529–2539.

Hu, L.; Gao, L.; Liu, Z.; and Feng, W. 2023c. Self-emphasizing network for continuous sign language recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 854–862.

Huang, J.; Zhou, W.; Zhang, Q.; Li, H.; and Li, W. 2018. Video-based sign language recognition without temporal segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Imashev, A.; Mukushev, M.; Kimmelman, V.; and Sandygulova, A. 2020. A Dataset for Linguistic Understanding, Visual Evaluation, and Recognition of Sign Languages: The K-RSL. In Fernández, R.; and Linzen, T., eds., *Proceedings of the 24th Conference on Computational Natural Language Learning*, 631–640. Online: Association for Computational Linguistics.

Jiao, P.; Min, Y.; Li, Y.; Wang, X.; Lei, L.; and Chen, X. 2023. CoSign: Exploring co-occurrence signals in skeleton-based continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20676–20686.

Jin, S.; Xu, L.; Xu, J.; Wang, C.; Liu, W.; Qian, C.; Ouyang, W.; and Luo, P. 2020. Whole-Body Human Pose Estimation in the Wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Joze, H. R. V.; and Koller, O. 2018. Ms-asl: A large-scale data set and benchmark for understanding american sign language. *arXiv preprint arXiv:1812.01053*.

Ko, S.-K.; Kim, C. J.; Jung, H.; and Cho, C. 2019. Neural sign language translation based on human keypoint estimation. *Applied sciences*, 9(13): 2683.

Koller, O.; Forster, J.; and Ney, H. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141: 108–125.

Koller, O.; Zargaran, S.; and Ney, H. 2017. Re-Sign: Re-Aligned End-to-End Sequence Modelling with Deep Recurrent CNN-HMMs. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3416–3424.

Koller, O.; Zargaran, S.; Ney, H.; and Bowden, R. 2016. Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition. In *BMVC*, 136–1.

Li, D.; Rodriguez, C.; Yu, X.; and Li, H. 2020a. Word-level deep sign language recognition from video: A new large-scale dataset

and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1459–1469.

Li, D.; Xu, C.; Yu, X.; Zhang, K.; Swift, B.; Suominen, H.; and Li, H. 2020b. Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. *Advances in Neural Information Processing Systems*, 33: 12034–12045.

Li, D.; Yu, X.; Xu, C.; Petersson, L.; and Li, H. 2020c. Transferring cross-domain knowledge for video sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6205–6214.

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

Liu, Y.; Gu, J.; Goyal, N.; Li, X.; Edunov, S.; Ghazvininejad, M.; Lewis, M.; and Zettlemoyer, L. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8: 726–742.

Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.-L.; Yong, M. G.; Lee, J.; et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Meng, Z.; Wu, Y.; Kanda, N.; Lu, L.; Chen, X.; Ye, G.; Sun, E.; Li, J.; and Gong, Y. 2021. Minimum word error rate training with language model fusion for end-to-end speech recognition. *arXiv preprint arXiv:2106.02302*.

Min, Y.; Hao, A.; Chai, X.; and Chen, X. 2021. Visual alignment constraint for continuous sign language recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11542–11551.

Neelakantan, A.; Xu, T.; Puri, R.; Radford, A.; Han, J. M.; Tworek, J.; Yuan, Q.; Tezak, N.; Kim, J. W.; Hallacy, C.; et al. 2022. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Papadimitriou, K.; and Potamianos, G. 2020. Multimodal Sign Language Recognition via Temporal Deformable Convolutional Sequence Learning. In *Interspeech*, 2752–2756.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Sincan, O. M.; and Keles, H. Y. 2020. AUTSL: A Large Scale Multi-Modal Turkish Sign Language Dataset and Baseline Methods. *IEEE Access*, 8: 181340–181355.

Sridhar, A.; Ganesan, R. G.; Kumar, P.; and Khapra, M. 2020. Include: A large scale dataset for indian sign language recognition. In *Proceedings of the 28th ACM international conference on multimedia*, 1366–1375.

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

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tunga, A.; Nuthalapati, S. V.; and Wachs, J. 2021. Pose-based sign language recognition using GCN and BERT. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 31–40.

von Agris, U.; Knorr, M.; and Kraiss, K.-F. 2008. The significance of facial features for automatic sign language recognition. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, 1–6.

Voskou, A.; Panousis, K. P.; Kosmopoulos, D.; Metaxas, D. N.; and Chatzis, S. 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.

Wang, H.; Chai, X.; Hong, X.; Zhao, G.; and Chen, X. 2016. Isolated sign language recognition with grassmann covariance matrices. *ACM Transactions on Accessible Computing (TACCESS)*, 8(4): 1–21.

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

Wong, R. C.; Camgöz, N. C.; and Bowden, R. 2024. SIGN2GPT: leveraging large language models for gloss-free sign language translation. In *ICLR 2024: The Twelfth International Conference on Learning Representations*.

Yang, Z.; Zeng, A.; Yuan, C.; and Li, Y. 2023. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4210–4220.

Yin, A.; Zhong, T.; Tang, L.; Jin, W.; Jin, T.; and Zhao, Z. 2023. Gloss attention for gloss-free sign language translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2551–2562.

Yin, K.; and Read, J. 2020. Better Sign Language Translation with STMC-Transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, 5975–5989.

Zhang, J.; Zhou, W.; Xie, C.; Pu, J.; and Li, H. 2016. Chinese sign language recognition with adaptive HMM. In *2016 IEEE international conference on multimedia and expo (ICME)*, 1–6. IEEE.

Zhang, W.; Huang, M.; Zhou, Y.; Zhang, J.; Yu, J.; Wang, J.; and Xu, L. 2024. BOTH2Hands: Inferring 3D Hands from Both Text Prompts and Body Dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2393–2404.

Zhao, R.; Zhang, L.; Fu, B.; Hu, C.; Su, J.; and Chen, Y. 2024. Conditional variational autoencoder for sign language translation with cross-modal alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19643–19651.

Zheng, J.; Wang, Y.; Tan, C.; Li, S.; Wang, G.; Xia, J.; Chen, Y.; and Li, S. Z. 2023. Cvt-slr: Contrastive visual-textual transformation for sign language recognition with variational alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 23141–23150.

Zhou, B.; Chen, Z.; Clapés, A.; Wan, J.; Liang, Y.; Escalera, S.; Lei, Z.; and Zhang, D. 2023. Gloss-free sign language translation: Improving from visual-language pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20871–20881.

Zhou, H.; Zhou, W.; Qi, W.; Pu, J.; and Li, H. 2021a. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1316–1325.

Zhou, H.; Zhou, W.; Zhou, Y.; and Li, H. 2021b. Spatial-temporal multi-cue network for sign language recognition and translation. *IEEE Transactions on Multimedia*, 24: 768–779.