Deep Understanding of Sign Language for Sign to Subtitle Alignment

Youngjoon Jang†, Jeongsoo Choi†, Junseok Ahn, and Joon Son Chung, *Member, IEEE*

***Abstract*—The objective of this work is to align asynchronous subtitles in sign language videos with limited labelled data. To achieve this goal, we propose a novel framework with the following contributions: (1) we leverage fundamental grammatical rules of British Sign Language (BSL) to pre-process the input subtitles, (2) we design a selective alignment (SA) loss to optimise the model for predicting the temporal location of signs only when the queried sign actually occurs in a scene, and (3) we conduct self-training with refined pseudo-labels which are more accurate than the heuristic audio-aligned labels. From this, our model not only better understands the correlation between the text and the signs, but also holds potential for application in the translation of sign languages, particularly in scenarios where manual labelling of large-scale sign data is impractical or challenging. Extensive experimental results demonstrate that our approach achieves state-of-the-art results, surpassing previous baselines by substantial margins in terms of both frame-level accuracy and F1-score. This highlights the effectiveness and practicality of our framework in advancing the field of sign language video alignment and translation. The code is available at https://github.com/art-jang/sign-to-subtitle.**

***Index Terms*—Sign language alignment, video-text alignment, language processing.**

I. Introduction

Sign language plays a crucial role in enabling communication for deaf individuals. To foster communication between hearing and deaf individuals, it is important to explore applications and technologies that can interpret linguistic meanings from sign language videos. With advancements in deep learning, recent works [1], [2], [3], [4], [5], [6] have demonstrated promising results in automatic sign language recognition and translation. However, these achievements have been restricted to environments where continuous signing is manually segmented into individual clips. A key challenge in scaling up such tasks is the difficulty in acquiring large-scale sign language training data consisting of pairs of sentence-level text labels and videos.

In response to this, recent research has utilised sign language broadcasted on TV, which provides videos with continuous signing alongside the transcriptions of the corresponding spoken audio, to construct large-scale datasets [7], [8]. While these datasets represent a significant improvement in scale compared to the previous benchmarks [9], [10], [11], their supervision on signed content is limited in two aspects. First, it is noisy because the presence of a word in the subtitle does not guarantee that the word is signed, and variations in signing may occur for the same subtitles. Second, it is weak because the audio content and the subtitles are not always temporally aligned with the signs in sign language. The weak and noisy supervision resulting from these issues ultimately leads to low performance in automatic sign language translation tasks [8], [7], [12].

Interest in automatic sign subtitle annotation has recently increased as a way to effectively utilise data collected from TV broadcasts. Previous works [12], [13], [14] have primarily concentrated on identifying sparse correspondence between keywords in the subtitles and the signs. To achieve dense alignment between the subtitles and the signs, a method for sign sentence boundary detection based on human body keypoints is proposed in [15]. Recently, [16] introduced the task of *aligning subtitles in sign language videos* and developed a baseline model, but no further research has been published on this task due to its inherent difficulty.

In this study, our objective is to develop a framework capable of aligning asynchronous subtitles in sign language videos with limited labelled data. The overall system architecture is visualised in Fig. 1. We first introduce a subtitle preprocessing technique that converts natural language sentences into sign language-like sentences, drawing upon fundamental grammatical rules of sign language. It is noteworthy that previous studies in sign language research have overlooked the substantial differences in grammatical structures between sign languages and spoken languages [17]. For instance, [16] extract text features using frozen BERT [18] model pre-trained on a large corpus of natural language data and [12] employ a straightforward approach of filtering out stop words from subtitles. We reflect the grammar system of sign language, which offers a simpler textual representation compared to natural language, enhancing alignment performance.

We also tackle the challenges posed by the weak and noisy supervision in datasets collected from TV broadcasts. To compensate for the noisy supervision caused by ambiguities between sign and subtitle, we introduce a selective alignment loss. This loss function is designed to optimise the model for predicting the precise temporal location of sign language occurrences within the scene, leveraging negative text-video pairs. By focusing on continuous frames where the sign language matches the queried subtitle, our model can avoid erroneous alignments that do not correspond to the intended text queries. Additionally, to mitigate the effect of weak supervision stemming from audio-aligned labels, we implement a self-training process. Specifically, we adopt a semi-

Y. Jang, J. Choi, J. Ahn, and J. S. Chung are with the School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), 291 Daehak-ro, Yuseong-gu, Daejeon, 34141, Republic of Korea (e-mail: wgs01088@kaist.ac.kr; jeongsoo.choi@kaist.ac.kr; junseok.ahn@kaist.ac.kr; joonson@kaist.ac.kr). Corresponding author: J. S. Chung. †Both authors contributed equally to this work.

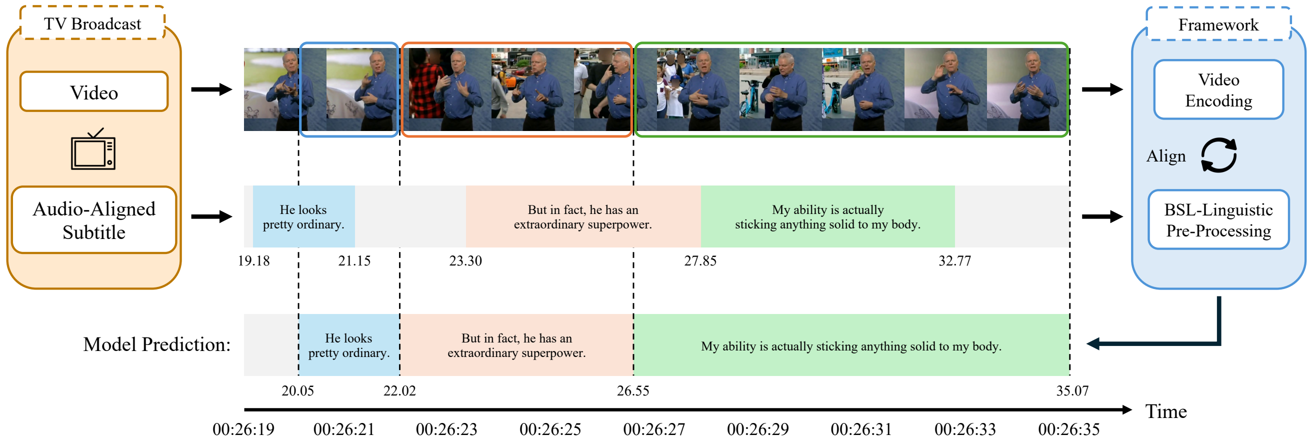

TV Broadcast
Video
Audio-Aligned Subtitle
Framework
Video Encoding
Align
BSL-Linguistic Pre-Processing
He looks pretty ordinary.
But in fact, he has an extraordinary superpower.
My ability is actually sticking anything solid to my body.
19.18
21.15
23.30
27.85
32.77
Model Prediction:
He looks pretty ordinary.
But in fact, he has an extraordinary superpower.
My ability is actually sticking anything solid to my body.
20.05
22.02
26.55
35.07
Time
00:26:19
00:26:21
00:26:23
00:26:25
00:26:27
00:26:29
00:26:31
00:26:33
00:26:35

Fig. 1: This work aims to align subtitles with continuous signing in sign language interpreted TV broadcast data by leveraging the grammatical systems of British Sign Language. Using two different modalities—video and audio-aligned subtitles—our framework encodes visual features and pre-processes the input query text based on the linguistics of BSL. The output consists of time segments that indicate the points in time when the sign language corresponding to the text is uttered.

supervised learning approach where the model is retrained using pseudo-labels generated by itself. This iterative process allows the model to refine its predictions by leveraging its own high-confidence outputs as additional supervision, thereby improving robustness under limited or noisy annotations. Our findings indicate that the accuracy of pseudo-labels generated by our model – trained on audio-aligned data and fine-tuned on manually labelled data – is 13.24% higher than that of audio-aligned labels in frame-level alignment accuracy (75.64% vs. 62.40%). As a result, we adopt a self-training strategy where the model is re-trained using the generated pseudo-labels. This iterative process leads to further improvement in model performance.

Our contributions can be summarised as follows: (1) to our best knowledge, we are the first to introduce a subtitle pre-processing algorithm aimed at trimming subtitles by incorporating the grammatical characteristics of sign language, (2) we propose a selective alignment (SA) loss, facilitating precise alignment with subtitle by preventing misalignment between unrelated video segment and text query, and (3) we employ a self-training strategy to address the weak supervision arising from heuristic audio-aligned subtitle data. With extensive experiments, we demonstrate that our proposed method achieves state-of-the-art results in both frame-level accuracy and F1 scores. These findings highlight the effectiveness and practical utility of our framework in advancing the field of sign language video alignment and translation.

II. Related Works

Sign language spotting. The sign language spotting task involves detecting isolated sign instances within continuous sign language videos. Early approaches use signing gloves [19] and hand-crafted visual elements to capture the hands, face, and motion, integrating these elements with Conditional Random Fields (CRFs) [20], [21], Hidden Markov Models (HMMs) [22], and Hierarchical Sequential Patterns (HSP) Trees [23]. Other approaches [24], [25] try to utilise subtitles of broadcast videos as auxiliary supervision to spot signs. With the advancement of deep learning, more recent work has leveraged extra cues such as mouthings [13] and visual dictionaries [14] or attached sliding window classifiers [26] to localise signs in time stamps more accurately. The most recent works [12], [27] focus on scaling up sign-spotting with large-scale corpora for automatic sign annotation. However, sign-spotting primarily focuses on detecting individual signs and does not address the broader alignment between video and sentence.

Continuous sign language recognition. The Continuous Sign Language Recognition (CSLR) task aims to map a given sign video to its corresponding gloss[1] sequence. Early works design hybrid models combining CNNs with HMMs [28], [29]. Later, Connectionist Temporal Classification (CTC) loss [30] is employed to facilitate training of CSLR models [31], [32], [33]. As a result, it allows the CSLR task to be considered as an alignment task between gloss sequence and sign video. From this, one branch focuses on designing model architectures [34], [35], [36], [37] specialised in temporal alignment, while the other branch [28], [29], [38], [39], [40] focuses on generating pseudo-gloss labels to propagate frame- or clip-level alignment supervision. However, existing methods have primarily been validated on pre-segmented sentences of signing [9], [41]. This is a major obstacle to applying CSLR technology in the wild, where sign language is continuously streamed.

Moment retrieval for video-text alignment. The Moment Retrieval (MR) task involves identifying temporal moments highly relevant to a given text query within a specified timeframe. This task is typically categorised into two approaches: proposal-based and proposal-free methods. Proposal-based methods use a pipeline where candidate windows are generated from the entire video, followed by ranking based on matched scores using predefined temporal structures like sliding windows [42], [43], [44], [45] or temporal anchors [46], [47], [48], [49], [50]. Proposal-free methods treat the task as a regression problem, directly regressing start and end time frames using multimodal attention, dynamic filters, and additional

[1] The smallest unit having independent meaning in sign language.

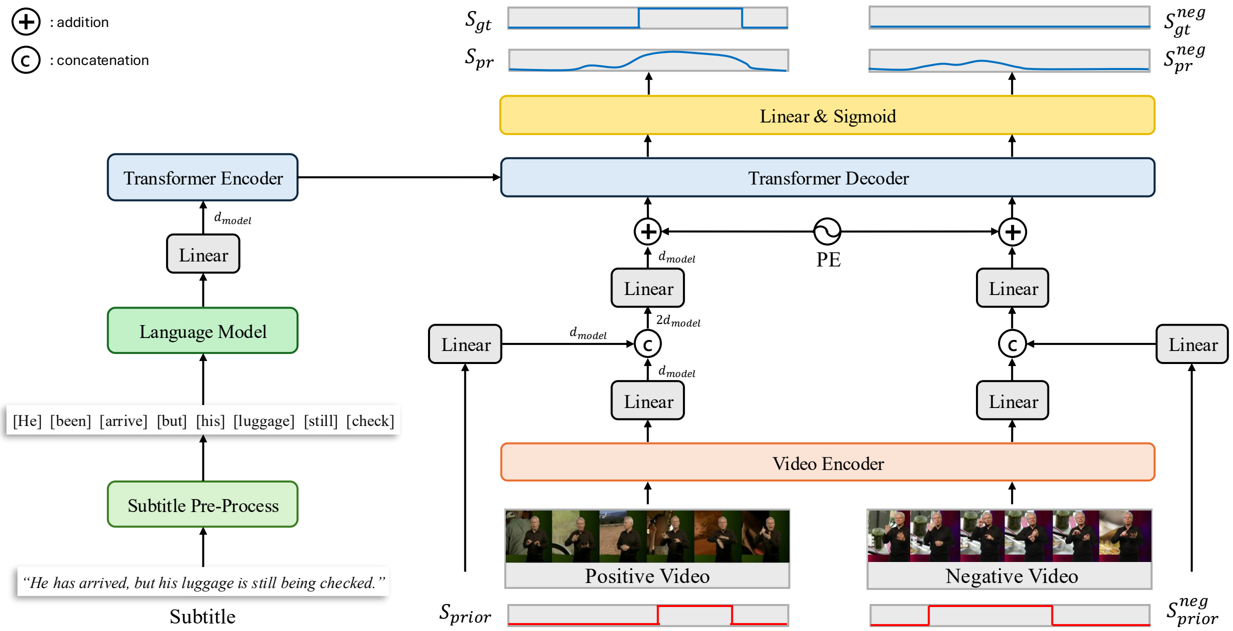


Fig. 2: **An illustration of our framework.** We input to our model: (1) a pseudo-gloss sequence pre-processed by a subtitle pre-processing mechanism, (2) a positive video aligned with the input query, and (3) a negative video not aligned with the input query. Both videos are encoded with shifted temporal boundaries of the audio-aligned subtitle, denoted as S_{prior} and S_{prior}^{neg}. Using a Transformer decoder, the model predicts frame-level alignment between text and video. Note that the negative video is only provided during training.

features [51], [52], [53], [54], [55], [56], [57]. MR shares similarities with sign language subtitle alignment, as both tasks involve aligning text and video at a frame level. However, there are distinct differences: (1) sign language content requires fine-grained alignment due to the consistent visual appearance of signing sequences across frames, necessitating precise recognition of body dynamics. (2) unlike other alignment tasks, each subtitle in sign language alignment has a specific reference location, which provides prior information about its start time and duration. In this paper, our primary focus is on the sign language subtitle alignment task, addressing the unique challenges and considerations inherent to this domain.

Different learning strategies of sign language. DNF [32] presents an iterative framework for CSLR, where an end-to-end model is first trained to obtain temporal alignments, which are then used to refine the feature extractor. SignBERT [58] proposes a self-supervised pre-training method using 3D hand reconstruction from 2D keypoints. This enables learning of rich structural features. SignBERT+ [59] extends this approach to multiple sign language tasks by slightly modifying the sequence model while maintaining the shared backbone, showing strong generalisability. SOS [33] tackles background bias via a teacher-student framework, where the student is trained on background-augmented videos and the teacher on originals, promoting background-invariant feature learning.

Alignment between audio, visual, and text modalities. Whisper [60] is a speech recognition model that performs both transcription and speech-text alignment. Its extension, WhisperX [61], further achieves word-level alignment, offering finer temporal precision. Multimodal alignment has also been studied in tasks such as speaker diarisation [62], [63], [64], which identifies who spoke what and when, combining audio, visual, and textual cues. Recent efforts [65], [66] include automating subtitle generation for film, requiring accurate video-text alignment. Beyond audio, DVFA [67] aligns talking face videos and transcriptions without using speech, achieving sentence-level alignment. EVFA [68] improves on this by leveraging global-local attention to capture both local details and global context, enhancing visual-text alignment. These works demonstrate the growing focus on aligning visual and textual information, even in the absence of audio.

Sign language subtitle alignment. Early work [69] on aligning subtitles to continuous signing combined cues from multiple sparse correspondences. However, this work was based on the assumption that the ordering of words in subtitles is the same as that of signing. Since then, the sequence-level sign language temporal localisation task has been studied using various approaches such as category-agnostic sign segmentation [70], [71], signer diarisation [72], [73], [74], and active signer detection [75], [76], [77], [78]. Nevertheless, these studies deal with temporal granularity, such as word boundaries or active sign segments, rather than subtitle units. In response to this, [15] initially introduce a model for segmenting sign language video into sentence-like units. While the model is accurate at detecting temporal boundaries, achieving an ROC-AUC metric of 0.87 at the frame level, aligning these segmented units with corresponding subtitles requires additional steps such as sign language translation [9], [79], [7], [2], [80], [37] and text processing. More recently, the sentence-level sign alignment task has been extended by simulating real-world data acquisition scenario [16] using a large-scale broadcast dataset [8]. However, other than this study, research on sign language subtitle alignment in real-life scenarios defined in [16] has not been actively conducted due to the high entry barrier posed by the difficulty of this task. In this paper, we present a novel task-specific framework that significantly outperforms the baseline's alignment performance with a linguistic analysis of British Sign Language.

III. Method

A. Sign to subtitle alignment framework

Our framework is illustrated in Fig. 2. For the input of the model, we provide (1) a pseudo-gloss sequence pre-processed by a subtitle pre-processing algorithm, (2) a positive video aligned with the input query, and (3) a negative video not aligned with the input query. Since guiding the model with prior which is a temporal boundary from the audio-aligned subtitle has a significant impact on model performance [16], both videos are encoded with the priors denoted as S_{prior} and S^{neg}_{prior}. Finally, the model produces a vector of values ranging from 0 to 1, indicating the relevance of each frame to the text query. The first and last values exceeding a threshold τ define the predicted temporal boundaries for the query subtitle. Notably, during inference, only the positive video is inputted.

Subtitle pre-processing. With an understanding of British Sign Language (BSL) linguistics, we transform the text query input into a sign language-like sentence. Further details about this process will be provided in Sec. III-B.

Text encoding. The pre-processed subtitle is encoded with a language model. Specifically, we initialise our language model using pre-trained BERT [18] weights, which were trained on the BookCorpus and English Wikipedia. The language model generates a sequence of contextualised token features, with special tokens representing the beginning and end of each sentence. To align with the input dimension of the Transformer encoder, these sequence features are projected to d_{model} using a linear layer. Subsequently, the output from the linear layer is fed into the Transformer [81] encoder to propagate token-level correlation information for the decoding process.

Video encoder. We input a video sequence of length T into the visual encoder for feature extraction. The encoded features are projected to d_{model} to be combined with the text modality in the Transformer decoder.

Prior encoding. Following the approach in [16], we utilise audio-aligned subtitle timings as a prior cue for subtitle localisation. Specifically, given the temporal boundary of an audio-aligned subtitle, denoted as S_{audio}, we shift the centre by $+2.7$ seconds to obtain an adjusted location S^{+}_{audio}. This temporal shift is motivated by observations in SAT [16], where shifting the prior boundary slightly forward in time was shown to improve alignment performance.

Unlike SAT, which uses the adjusted location S^{+}_{audio} directly as a narrow prior, we further extend this range by spanning 3.2 seconds (i.e., 20 frames) on both sides, resulting in a broader prior region, denoted as S_{prior}. This design encourages the model to avoid overly focusing on a narrow segment, which can lead to misalignment—particularly for short subtitles, where the correspondence between audio and subtitle is often less precise. The effectiveness of this broader prior is demonstrated by the performance improvement shown in Table VII.

The prior S_{prior} is encoded as a binary vector of length T, where each element is set to 1 if the corresponding video frame falls within the temporal span of S_{prior}, and 0 otherwise. This binary prior is first projected to a feature dimension of d_{model}. Since the video features are also of dimension d_{model}, the two are concatenated along the channel axis, resulting in a combined feature of dimension $2d_{\text{model}}$. This concatenated representation is then projected back to d_{model}, allowing the model to effectively incorporate the temporal prior into the video feature representation.

Multimodal decoding. The decoder comprises Transformer layers that process the encoded sequence to aggregate visual and text features. Positional encoding is applied to capture the temporal order of signing. The final layer is a linear transformation followed by a sigmoid activation, producing T predictions between $[0, 1]$ for each video frame.

B. Subtitle pre-processing

The grammatical system of British Sign Language (BSL) significantly differs from that of spoken languages [17]. However, existing deep learning-based sign language-related studies simply adopt techniques from the Natural Language Processing (NLP) field, such as freezing pre-trained language models or reducing noise in the input text. In this study, we explore the impact of sign language grammar rules on alignment performance and propose an online subtitle pre-processing algorithm. This algorithm transforms natural language sentences into sentences that mimic the grammar system of sign language, in other words, resemble gloss representation. We delve into the grammar of BSL based on insights from [17]. We define guidelines that characterise the unique linguistic structure of BSL: (1) pronouns extend beyond referring to people and also include concepts of places where people exist, (2) negative meaning is conveyed through negative phrases, (3) articles may be omitted in gloss representation, (4) tense may not always be explicitly indicated, (5) 'be' verbs may be omitted in gloss representation, and (6) 'been' word is used to express the present perfect tense. Reflecting on these rules, the following strategies are adopted: (1-2) we do not remove stopwords that are considered an unimportant component in NLP field, (3) exclude articles, (4) lemmatise verbs (5) eliminate 'be' verbs, and (6) replace the auxiliary verb colour orange'have' with 'been'. At the last stage, we analyse the role of the word 'have' in a sentence using the Part of Speech (PoS) tagging approach. Our experiments prove that tailoring input queries based on the grammatical rules of BSL significantly enhances alignment performance. This result underscores the importance of incorporating sign language-specific linguistic features in the subtitle alignment domain.

C. Training and fine-tuning

The sign language subtitle alignment task ultimately aims to achieve auto-labelling while minimising the reliance on extensive human labelling. In this study, we simulate a scenario where the amount of manually labelled data is significantly less compared to the volume of audio-aligned labelled data. This approach reflects real-world challenges where acquiring large-scale manually labelled datasets for sign language alignment is resource-intensive and impractical.

Our training strategies are composed of three stages. Firstly, we perform word pre-training, a task involving single sign-spotting under the same setup of [16]. Notably, our model can be applied to the sign-spotting task without any modification, except for the input data type (sentence → word), achieving improved performance compared to the previous method. Experimental results are detailed in Sec. IV-B. Secondly, we conduct a training stage using weak labels, specifically audio-aligned subtitles. This stage focuses on helping the model learn from various text corpus. Finally, to refine model predictions, we fine-tune the model using a small amount of manually labelled subtitle data. During these phases, the model is trained with a binary cross-entropy loss between the predicted vector S_{pr} and the ground truth S_{gt}, defined as follows:

$$\mathcal{L}_{align} = -\frac{1}{T}\sum_{t=1}^{T}\left(S_{gt,t}\log S_{pr,t} + (1 - S_{gt,t})\log(1 - S_{pr,t})\right), \tag{1}$$

where t denotes a frame index. This loss encourages our model to discriminate which frames are relevant to the subtitle and which are not.

D. Selective alignment loss

As explained in Sec. III-C, our model is trained on audio-aligned subtitles in the training stage. In practice, S_{prior} is the timing of the temporally jittered audio-aligned subtitle. However, this setup leads to increased dependence on prior information. To solve this problem, this paper introduces selective alignment (SA) loss, which is composed of negative alignment loss and relative alignment loss.

Negative alignment loss. To reduce dependence on the prior, we introduce negative video-subtitle pairs to simulate scenarios where our model should predict non-aligned frames based on text understanding. In our implementation, when the model receives a negative pair as input, the ideal output should be a zero-valued sequence vector because there is no video frame associated with the given subtitle. We use binary cross-entropy loss to train the model with a zero-valued sequence vector as the label for negative video pairs. The negative alignment loss can be formulated as follows:

$$\mathcal{L}_{neg} = -\frac{1}{T}\sum_{t=1}^{T}\log(1 - S^{neg}_{pr,t}), \tag{2}$$

where S^{neg}_{pr} represents the model prediction when the negative pair is given as the input. This loss function helps the model learn to distinguish misaligned video-subtitle pairs more effectively, thereby improving its alignment capabilities.

Relative alignment loss. When negative samples are given into the model during training, a bias arises where the model tends to output more zero-valued sequences. To address this issue, inspired by [82], we propose a relative alignment loss that leverages the provided ground truth labels to differentiate between probabilities of aligned frames from those of non-aligned frames. To compute the relative importance, the proposed loss is defined as follows:

$$\mathcal{L}_{rel} = -\frac{1}{\sum_{t=1}^{T} S_{gt,t}}\sum_{t=1}^{T} S_{gt,t}\log\frac{\mathrm{e}^{S_{pr,t}}}{\sum_{i=1}^{T}\mathrm{e}^{S_{pr,i}} + \mathrm{e}^{S^{neg}_{pr,i}}}. \tag{3}$$

This loss function aims to mitigate the aforementioned challenges and encourage the model to predict alignments more effectively, reducing sensitivity to the number of aligned frames in the ground truth. The total loss $\mathcal{L}_{tot}$ is as follows:

$$\mathcal{L}_{tot} = \mathcal{L}_{align} + \lambda_{neg}\mathcal{L}_{neg} + \lambda_{rel}\mathcal{L}_{rel}, \tag{4}$$

where λ_{neg} and λ_{rel} are hyperparameters for negative alignment loss and relative alignment loss, respectively. In our experiments, λ_{neg} and λ_{rel}, are set to 1. Note that the model empirically shows robust performance across a range of values between 0.5 and 2.0, allowing us to fix both hyperparameters to 1 throughout.

E. Semi-supervised data annotation

In this study, our model is trained using audio-aligned subtitle temporal boundaries, which can be obtained relatively easily, and fine-tuned with a small amount of human-labelled data. Although fine-tuning improves alignment performance, we revisit the limitations introduced by the weak supervision of the initial training stage. To address this issue, we employ self-training, a widely used semi-supervised learning approach in which a fully trained model generates pseudo-labels for unlabelled data and is subsequently retrained on this automatically labelled data.

Our self-training framework for the alignment task consists of two main steps. First, we compute frame-level alignment confidence scores from pairs of sign language video clips and text queries. Then, the peak confidence score is compared against a predefined threshold τ_c to determine whether the sample should be included in the self-training set. Samples with peak scores below the threshold are filtered out to ensure pseudo-label reliability.

In addition, we analyse the relationship between data quality and quantity by varying the confidence threshold τ_c. Experimental results demonstrate that self-training effectively enhances model performance, highlighting its usefulness under weak supervision.

IV. Experiments

A. Experimental setup

Dataset. Our framework is trained on the BOBSL dataset [8], a publicly available dataset comprising British Sign Language interpreted BBC broadcast footage, accompanied by English subtitles corresponding to the audio content. This dataset consists of 1,940 episodes totaling 1,447 hours, with 1.2M sentences covering a vocabulary of 77K words. It involves 37 signers. The experimental setup, including the dataset split, follows the publicly available code of [16] provided

TABLE I: **Performance comparison on BOBSL dataset.** Our method outperforms existing approaches across all metrics, demonstrating its capability to align subtitles with sign videos.

Method	frame-acc (%) ↑	F1@.10 ↑	F1@.25 ↑	F1@.50 ↑
S_{audio}	40.14	46.36	33.46	14.10
S^{+}_{audio}	62.40	72.79	64.09	44.60
SAT [16]	71.01	74.19	67.01	53.36
DVFA [67]	72.55	74.87	67.70	55.63
Ours	**77.22**	**81.39**	**75.03**	**63.81**

TABLE II: **Evaluation on downstream tasks.** The pseudo-subtitles generated by our framework contribute to performance improvement in both retrieval and continuous sign language recognition (CSLR) tasks.

Task	Retrieval				CSLR
Metric	T2V		V2T		WER ↓
	R@1↑	R@5↑	R@1↑	R@5↑	
$CSLR^2$ [84] w/SAT [16]	27.14	42.19	26.25	41.99	65.16
$CSLR^2$ [84] w/Ours	**28.59**	**43.74**	**26.59**	**42.51**	**64.22**

by their official repository[2]. We use a subset of this dataset, specifically 16 episodes, for fine-tuning our model and 35 episodes for testing. The test set encompasses 30 hours of video and includes 20,338 English subtitles, with a total vocabulary of 13K words. In this research, we hypothesise that the cost of manually labelling alignment data can be minimised by leveraging a small amount of human-labelled data, representing approximately 2% of the volume of heuristic audio-aligned data available. This approach demonstrates our strategy to efficiently utilise limited labelled data to achieve state-of-the-art performance in sign language alignment tasks.

To further validate the generalisability of our model, we use the SRF subset of the WMT-SLT dataset [83], which consists of daily national news and weather forecast episodes broadcast by Swiss national TV and narrated in standard Swiss German. The dataset contains 29 episodes with a total duration of 16 hours. We randomly select two episodes, assigning one to validation and the other to testing.

Evaluation protocol. We follow the evaluation protocol proposed in [16], which utilise frame-level accuracy and F1-score to assess the performance of alignment models. For evaluating the F1-score, we quantify the accuracy of subtitle alignment with sign language videos by considering hits and misses based on temporal overlaps. Specifically, we use three temporal overlap thresholds (IoU $\in$ $\{0.1, 0.25, 0.50\}$, representing F1@.10, F1@.25, and F1@.50, respectively). These metrics compare the predicted spans S$_{pr}$ with ground truth subtitle spans S_{gt} to measure alignment accuracy. The frame-level accuracy is denoted as frame-acc in our tables. For evaluating the sign-spotting task, we calculate two metrics: mean average precision (mAP) and top-1 classification accuracy (Acc@1).

Implementation details. In subtitle pre-processing, we utilise the pycontractions library[3] for fixing contractions and the Natural Language Toolkit [85] for lemmatisation and Part of Speech (PoS) tagging. For model architecture, we follow the architecture of SAT model. The language model is based on the BERT architecture, featuring 12 layers, 12 attention heads, and a model size of 768. We use a vanilla 2-layer Transformer encoder and decoder with 4 attention heads, where the Transformer's model dimension d_{model} is set to 512. The video encoder leverages a pre-trained I3D [86] sign classification model [12], producing 1024-dimensional visual embeddings pre-extracted from sign language video segments. For positional encodings in the input to the video encoder, we employ 512-dimensional sinusoidal positional encodings. During training, we randomly select a 20-second search window around the location of the ground truth subtitle S_{gt}. Within this window, we sample video features with a stride of 4. Note that S_{gt} represents an audio-aligned subtitle during training and transitions to a sign-aligned subtitle during fine-tuning. To ensure a fair comparison with previous methods, we use a DTW threshold of 0.4 for global alignment in long-sequence videos. We use the Adam [87] optimiser with a batch size of 64 for training. During the word pre-training stage, the learning rate is set to 10^{-5}, and for training and fine-tuning with subtitles, we use a learning rate of 5×10^{-6}. The word pre-training stage involves training the model for 7 epochs, followed by 4 epochs of training for subsequent stages. Full-sentence fine-tuning is conducted over 100 epochs.

B. Quantitative results

Comparison to baselines on BOBSL dataset. As shown in Table I, first, we evaluate two baseline approaches: using the original, non-shifted audio-aligned subtitle S_{audio} and using the shifted audio-aligned subtitle S^{+}_{audio} (shifted by 2.7 seconds). The results demonstrate significant performance improvements by simply shifting the time of the audio-aligned subtitle. Next, we assess the performance of the SAT [16] model, which represents the initial work in this field. The SAT model outperforms all baselines by leveraging subtitle text to identify associated video segments. Although SAT is the only prior work that explicitly performs text-based sign-to-subtitle alignment from audio-aligned subtitles, we further explore a stronger comparison by adapting DVFA [67], a recent model originally designed for aligning talking face videos with transcriptions. For a fair comparison, we replace the talking face video with sign language video and use the subtitle as input transcription. All training hyperparameters are unified with our training pipeline. As a more recent model, DVFA achieves better performance than SAT, highlighting the benefits of deeper video-text alignment modelling. Finally, our proposed method exhibits superior performance across all metrics by a substantial margin. This underscores our framework's capability to align subtitles with sign videos based on a deep understanding of BSL's linguistic characteristics. Note that all metric is calculated with Dynamic Time Warping (DTW) for global alignment in long-time sequence videos with overlapped regions removed, following the method described in [16].

Evaluation on downstream tasks, sign language retrieval and continuous sign language recognition. For the baseline of both retrieval and Continuous Sign Language Recognition

[2] https://github.com/hannahbull/subtitle_align

[3] https://pypi.org/project/pycontractions

TABLE III: **Sign word spotting performance.** We randomly jitter prior S_{prior}. Our model shows superior performance in the sign spotting task.

Method	mAP ↑	Acc@1 (%) ↑
Random prior	7.60	4.67
SAT [16]	70.83	68.94
Ours	**76.89**	**75.35**

TABLE IV: **Comparison with gloss.** We emphasise that our subtitle pre-processing does not require any training stage or manual labelling process, unlike gloss production.

Input	frame-acc (%) ↑	F1@.10 ↑	F1@.25 ↑	F1@.50 ↑
Pseudo-gloss	72.92	78.43	71.10	57.12
Stemming	74.66	79.01	71.99	59.35
Subtitle pre-process (**Ours**)	**75.64**	**80.02**	**73.35**	**61.45**

(CSLR) tasks, we utilise the $CSLR^2$ model [84], which is trained using pseudo-subtitle labels generated by the SAT model. We train the $CSLR^2$ model from scratch with pseudo-subtitles automatically generated by our pipeline, which serve as sentence-level training data within the BOBSL dataset.

Since BOBSL does not provide manually aligned sentence-level annotations for training, these pseudo-subtitles play a crucial role in creating sentence-level training data, directly impacting the quality of the learning signal. For evaluation, we use the CSLR-TEST subset of BOBSL, which contains 4,590 sentences, where BSL signing sequences are manually aligned with their corresponding English subtitle sentences.

For retrieval task, we report both text-to-video (T2V) and video-to-text (V2T) retrieval results using recall at rank k (R@k) for $k \in \{1, 5\}$. For CSLR, we report word error rate (WER), which quantifies transcription accuracy by calculating the number of substitutions, deletions, and insertions needed to convert the predicted transcription into the ground truth, normalised by the total number of words in the ground truth.

As shown in Table II, the higher quality pseudo-subtitles generated by our method contribute to improved performance across all metrics. This demonstrates that our approach can provide more reliable sentence-level annotations, making it highly applicable to various sign language downstream tasks that require robust pseudo-subtitle training data.

Sign-spotting performance. As mentioned earlier, our model can be directly applied to the sign-spotting task without modifying the model architecture. To use the model in this way, we input a random prior into the model during both the training and testing stages. The sign-spotting task aims to localise the temporal position of a sign word given a textual query, which requires fine-grained alignment between sign sequences and individual words. To evaluate the model, we use the benchmark from [88], which provides temporal annotations of sign-word alignments obtained through mouthings and dictionary exemplars.

As shown in Table III, the random prior achieves a very low mean Average Precision (mAP) of 7.6. In contrast, the SAT model achieves a significantly higher mAP of 70.83 and an accuracy at rank 1 (Acc@1) of 68.94%. Acc@1 measures the proportion of samples in which the top-ranked prediction exactly matches the ground truth temporal segment. Our model outperforms both baselines, achieving an mAP of 76.89 and Acc@1 of 75.35%.

These results highlight the effectiveness and generalisability of our model to downstream tasks requiring a detailed understanding of sign language, such as sign spotting.

Comparison with gloss representation. To verify whether the proposed subtitle pre-processing serves as an effective representation for the sign-to-subtitle alignment task, we compare it against a linguistically motivated alternative: gloss-level supervision and stemming. Glosses represent the minimal meaningful units in sign language, capturing its grammar and structure more explicitly than raw subtitles. However, for British Sign Language (BSL), there exists no publicly available ground-truth gloss annotation or reliable text-to-gloss conversion model. Therefore, we adopt pseudo-glosses automatically extracted by [88], which offer an approximate gloss-level representation.

In this experiment, we replace our subtitle pre-processing pipeline with two alternative input representations—pre-extracted pseudo-glosses and stemmed subtitles—while keeping all other training conditions identical. This allows us to isolate the effect of the input representation on alignment performance. As shown in Table IV, using pseudo-glosses yields a frame-level accuracy of 72.92%, while applying stemming improves it slightly to 74.66%. Our proposed subtitle pre-processing further increases the accuracy to 75.64%, outperforming both alternatives. These results indicate that our approach not only preserves the semantic content necessary for alignment but also provides a stronger learning signal for capturing the underlying sign language structure. We attribute this improvement to the fact that our subtitle pre-processing retains contextual and grammatical cues that are often lost or distorted in pseudo-gloss extraction or by simple stemming.

Importantly, unlike gloss-based approaches that require dedicated models or annotation pipelines for gloss generation, our subtitle pre-processing method does not involve any learning or model training. This makes it significantly more efficient and scalable, as it can be directly applied to raw subtitle data without the need for additional supervision or resources. Note that we do not apply self-training to either of the two models in this comparison: one trained with subtitle pre-processing and the other with pseudo-glosses. This ensures that any performance difference arises purely from the representational quality of the input, rather than differences in training strategy.

Analysis of input data at inference. We create subsets from the BOBSL test set through various filtering strategies and explore the corresponding performance variations, as shown in Table V. When evaluating only the data overlapping with S_{prior} (Overlap only case), there is a significant performance boost of 12.35 in F1@0.50. This demonstrates that a better prior can contribute to improved performance and highlights the importance of properly setting the prior in the training data. Furthermore, when removing cases with insufficient language information—specifically, subtitles with fewer words (Remove shorter (n) cases)—we observe an overall improvement in

TABLE V: **Analysis of input data at inference.** *All*: the entire dataset; *Overlap only*: cases where there is an overlap between S_{prior} and the subtitle; *Remove shorter* (n): excludes subtitles with fewer than n words; *Remove longer* (n): removes subtitles with more than n words.

Data	#sent	frame-acc (%) ↑	F1@.10 ↑	F1@.25 ↑	F1@.50 ↑
All	20,338	77.22	81.39	75.03	63.81
Overlap only	15,898	80.80	91.27	86.81	76.16
Remove shorter (2)	19,204	77.55	83.71	77.95	66.64
Remove shorter (3)	17,817	77.95	85.44	80.25	68.96
Remove shorter (4)	16,406	78.48	87.02	82.44	71.22
Remove longer (50)	20,280	77.35	81.30	74.89	63.59

TABLE VI: **Performance comparison to baselines on WMT-SLT dataset.** Our method outperforms existing approaches across all metrics. Ours* denotes the DVFA model combined with the proposed components applied.

Method	frame-acc (%) ↑	F1@.10 ↑	F1@.25 ↑	F1@.50 ↑
Random prior	49.68	69.38	55.04	28.68
S^{+}_{audio}	71.38	85.77	85.77	85.77
SAT [16]	Failed to train			
DVFA [67] w/ Random prior	73.03	79.34	73.80	61.99
DVFA [67] w/ S^{+}_{audio}	79.14	93.26	91.33	87.11
Ours* w/ Random prior	77.22	83.03	79.70	67.53
Ours* w/ S^{+}_{audio}	**82.07**	**97.79**	**95.20**	**89.30**

performance. This indicates that a lack of linguistic content negatively affects model accuracy. Lastly, when removing cases with longer subtitles (Remove longer (n) cases), performance slightly decreases, suggesting that an excessive amount of language information does not necessarily lead to further improvements.

Comparison to baselines on WMT-SLT dataset. To demonstrate the generalisability of our pipeline, we evaluate our model on the SRF subset of the WMT-SLT dataset [83], which consists of approximately 16 hours of Swiss German news and weather broadcast data. We compare two types of priors: (1) random priors and (2) audio-aligned subtitle timings shifted by one second (denoted as S^{+}_{audio}). As shown in Table VI, S^{+}_{audio} achieves an F1 score of 85.77 across all IoU thresholds (0.1, 0.25, 0.5), suggesting that this prior is more accurate than those obtained from the BOBSL dataset.

The SAT model fails to converge on the SRF subset, which indicates that the vanilla Transformer architecture requires more data to learn effectively. In contrast, DVFA, based on a Conformer [89], successfully learns even with the limited dataset. When using random priors, DVFA achieves a frame-level accuracy of 73.03%, which further improves to 79.14% with the stronger S^{+}_{audio} prior, confirming that better priors lead to better alignment. Furthermore, we enhance DVFA by introducing (1) subtitle pre-processing and (2) the SA loss during training, denoted as Ours*. These components consistently improve performance, and when combined with the S^{+}_{audio} prior, Ours* achieves a frame accuracy of 82.07%, demonstrating the effectiveness of the proposed approach.

TABLE VII: **Ablation study on model component.** we demonstrate that each element – subtitle pre-processing, selective alignment loss, and self-training – consistently contributes to performance improvements.

Method	frame-acc (%) ↑	F1@.10 ↑	F1@.25 ↑	F1@.50 ↑
SAT [16]	71.01	74.19	67.01	53.36
Baseline (SAT + span prior)	72.59 (+1.58)	77.04 (+2.85)	69.84 (+2.81)	56.24 (+2.88)
+ subtitle pre-processing	74.46 (+3.45)	78.62 (+4.43)	71.75 (+4.74)	59.50 (+6.14)
+ SA loss	75.64 (+4.63)	80.02 (+5.83)	73.35 (+6.34)	61.45 (+8.09)
+ self-training **(Ours)**	**77.22** (+6.21)	**81.39** (+7.20)	**75.03** (+8.02)	**63.81** (+10.45)

TABLE VIII: **Ablation study on effectiveness of each process in subtitle pre-processing.** This shows incremental performance improvements with each added process, achieving the best performance when all processes are applied.

Method	frame-acc (%) ↑	F1@.10 ↑	F1@.25 ↑	F1@.50 ↑
Baseline with SA loss	73.77	76.93	69.95	57.36
+ stopwords	75.23	79.57	73.02	60.72
+ remove articles	75.43	79.72	73.45	60.85
+ lemmatise	75.46	79.88	73.48	61.21
+ 'be' verb process **(Ours)**	75.64	80.02	73.35	61.45

For implementation details, we use a Video-Swin Transformer [90] pre-trained on the ISLR task using the BOBSL dataset (as in CSLR2 [84]) as our video encoder to extract visual features. The language encoder is initialised with pre-trained BERT weights, which are trained on a German corpus. Since the prior is randomly set during training, we do not perform prior span selection. For subtitle pre-processing, we follow the grammatical structure of German Sign Language (DGS), performing lemmatisation on verbs without removing stopwords.

These results demonstrate that the proposed components can be easily integrated into different model architectures and generalise well across languages.

C. Ablation study

Effectiveness of model component. We verify the effectiveness of the four components comprising our framework—(1) spanning prior, (2) subtitle pre-processing, (3) selective alignment (SA) loss, and (4) self-training—through various ablation studies. As reported in Table VII, alignment performance improves when introducing a spanning prior, where the subtitle boundary is extended by 3.2 seconds on both sides. This broader prior enables the model to learn more robustly from noisy audio-aligned timings by reducing sensitivity to temporal misalignment. Further improvements are observed when utilising subtitle pre-processing that incorporates the linguistic characteristics of sign language, enhancing both frame-level accuracy and F1 scores. Additionally, the inclusion of SA loss provides further gains by facilitating the learning of correlations between text and video using negative pairs. Finally, alignment is notably enhanced through a self-training process, where the model is refined using its own predictions instead of relying solely on weakly labelled audio-aligned subtitles. This analysis demonstrates that each component of our model contributes meaningfully to overall performance improvement.

Effectiveness of subtitle pre-processing. In Table VIII, we investigate the impact of reflecting sign language grammar in subtitle pre-processing on the model's understanding of

TABLE IX: **Ablation results according to the fine-tuning strategies.** Our final model freezes both Language Model and Transformer Encoder during the fine-tuning stage to prevent overfitting to a limited amount of labelled data.

Freeze					
Language Model	Transformer Encoder	frame-acc (%) ↑	F1@.10 ↑	F1@.25 ↑	F1@.50 ↑
		75.04	79.01	72.59	60.58
✓		**75.70**	79.93	73.26	61.37
✓	✓	75.64	**80.02**	**73.35**	**61.45**

TABLE X: **Ablation results according to the amount of fine-tuning data.**

Fine-tuning data ratio	frame-acc (%) ↑	F1@.10 ↑	F1@.25 ↑	F1@.50 ↑
0.0%	68.93	68.23	61.11	47.61
0.5%	74.37	78.23	71.60	59.46
1.0%	74.84	78.93	72.56	60.21
1.5%	75.53	79.51	73.18	61.06
2.0%	75.64	80.02	73.35	61.45

sign language. Instead of removing stopwords (including fixed contractions), we observe a notable performance improvement of 1.46% compared to the baseline. This finding suggests that stopwords carry significant meaning in sign language, unlike in traditional NLP tasks. The removal of articles and lemmatisation of verbs also contribute to a steady improvement in performance, indicating the importance of these pre-processing steps in enhancing the model's comprehension of signed language. The process of removing the 'be' verb and incorporating 'been' proves beneficial for the model's understanding of sign language. This adjustment improves the model's ability to capture sign language grammar and nuances effectively. Overall, these findings highlight the importance of tailored subtitle pre-processing techniques that align with the grammar and structure of signed language. Optimising these pre-processing steps can significantly enhance the model's performance and understanding of sign language.

Fine-tuning of subtitle encoder. In this paper, we explore the grammatical and linguistic system of BSL. This results in a unique form of text input that significantly differs from natural language. Therefore, we highlight the necessity of going beyond simply using conventional models employed in the NLP field during the text encoding process. Instead, we emphasise the importance of developing a sign language expert model, tailored to handle the intricacies of sign language linguistics.

To address this, we initialise our language model using pre-trained BERT [18] weights, which are trained on BookCorpus and English Wikipedia. However, unlike these conventional corpora, our subtitle inputs are linguistically distinct due to pre-processing tailored to sign language. To help the model adapt to these unique characteristics, we train both the BERT-based language model and the Transformer encoder during the pre-training stage, where audio-aligned subtitles—although weakly labelled—are relatively easy to obtain and cover a large, diverse corpus. This stage allows the model to learn a broader vocabulary and better accommodate the sign-specific linguistic structure of the pre-processed subtitles.

TABLE XI: **Ablation results according to the amount of the data used in the self-training stage.** The model's performance steadily improves with more data, which indicates that the output of our model is reliable.

Confidence threshold	Data ratio	frame-acc (%) ↑	F1@.10 ↑	F1@.25 ↑	F1@.50 ↑
Train with S^{+}_{audio}		75.50	79.73	73.29	61.29
0	100.0%	**77.22**	**81.39**	**75.03**	**63.81**
0.5	88.0%	77.09	80.91	74.70	63.51
0.9	62.2%	76.94	80.76	74.54	63.46
0.95	49.2%	76.60	80.11	73.96	62.78

TABLE XII: **Ablation study on selective alignment (SA) loss.** The best performance is achieved when both negative alignment loss $\mathcal{L}_{neg}$ and relative alignment loss $\mathcal{L}_{rel}$ are applied.

$\mathcal{L}_{neg}$	$\mathcal{L}_{rel}$	frame-acc (%) ↑	F1@.10 ↑	F1@.25 ↑	F1@.50 ↑
		74.46	78.62	71.75	59.50
	✓	74.32	78.34	71.75	59.36
✓		74.94	78.23	71.49	59.19
✓	✓	**75.64**	**80.02**	**73.35**	**61.45**

In contrast, during the fine-tuning stage, where the training data is fully labelled but significantly smaller, we freeze the language model and Transformer encoder to prevent overfitting and preserve the generalisable representations learned earlier. We conduct ablation studies to validate this strategy, as shown in Table IX. Training both the language model and Transformer encoder during fine-tuning leads to performance degradation, likely due to overfitting. Interestingly, freezing both modules yields similar performance to freezing only the language model, further supporting our approach.

This result demonstrates that training the model to learn a diverse vocabulary using large-scale audio-aligned data, while subsequently preserving it through freezing during fine-tuning, is an effective strategy for achieving generalisable and robust alignment performance.

Ablation results according to the amount of fine-tuning data. In Table X, we report the performance variation with respect to the amount of labelled data used during the fine-tuning stage. The fine-tuning data ratio indicates the proportion of manually aligned labels relative to the entire training set. A 0% ratio corresponds to a model without any fine-tuning, and we experiment with up to 2% of labelled data. As shown in the results, performance consistently improves as more human-labelled data is provided.

Although we are not able to obtain more than 2% of labelled data, we believe that further performance gains are likely with additional supervision. Notably, our framework achieves a significant result even with only 0.5% of labelled data—surpassing the performance of the SAT model trained with the full 2% labelled data, achieving a higher frame-level accuracy (74.37% vs 71.01%). This highlights the effectiveness and efficiency of our framework under limited supervision.

Ablation study on the amount of data used in self-training. We demonstrate the reliability of the fine-tuned model's outputs used in self-training, as detailed in Table XI.

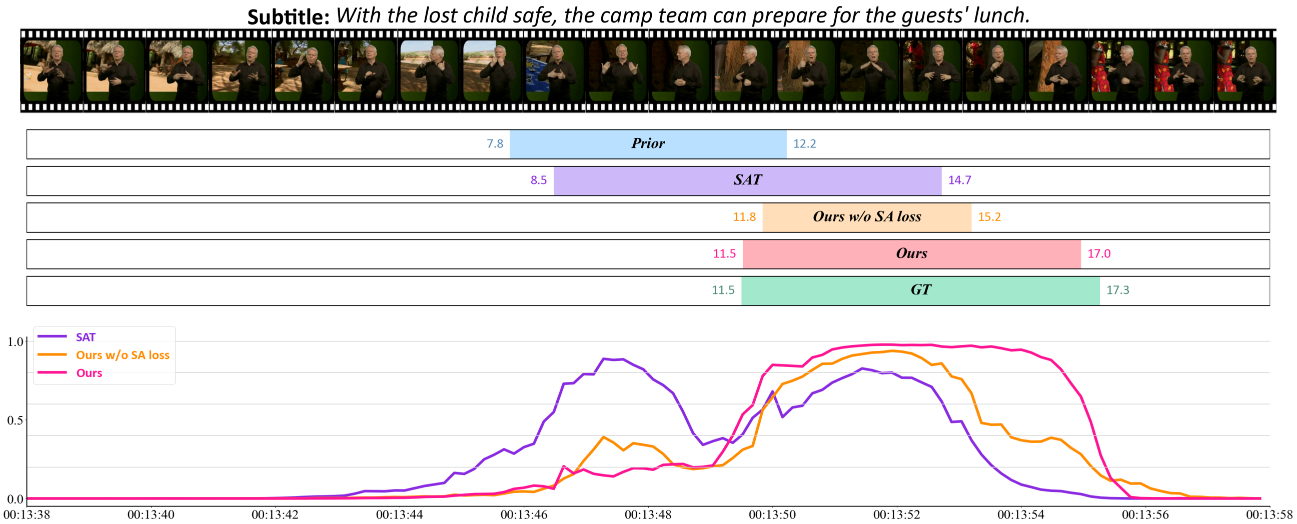


Fig. 3: Qualitative results. This figure presents frame-level model predictions, showcasing the timeline of prior, the SAT model, our model trained without the selective alignment (SA) loss, our full model, and ground truth. Our model, which incorporates rich sign language linguistics, shows superior alignment with the ground truth compared to other baselines. In addition, the frame-level alignment probability visualisation at the bottom of this figure illustrates our model's clear timeline alignment, demonstrating its effective correlation of text with video content.

To analyze the trade-off between the quality and quantity of pseudo-labels, we systematically vary the confidence threshold. Only pseudo-alignment labels exceeding the threshold are integrated into the self-training process. When the threshold is set to 0, all pseudo-labels generated on the training set are included. Setting the threshold to 0.5 results in approximately 88% of the training data being used, while thresholds of 0.9 and 0.95 reduce the included data to 62.2% and 49.2%, respectively. This setup allows us to examine how reducing the number of pseudo-labeled examples—by increasing the threshold—affects overall performance. The detailed procedure is described in Alg. 1.

Re-training the model with heuristic audio-aligned labels (denoted as 'Train with S^{+}_{audio}') under the same setting with self-training does not result in performance improvement (refer to the performance of 'Baseline + subtitle pre-processing + SA loss' in Table VII). By systematically varying the confidence threshold, we observe that self-training performance consistently improves as the amount of pseudo-labeled data increases, demonstrating that even lower-confidence labels generated by our model are reliable and beneficial. These results highlight that the pseudo-labels produced by our model, even at modest confidence levels, are of higher quality than heuristic audio-aligned subtitles, providing a stronger signal for effective self-training.

Algorithm 1 Self-training with pseudo-labelling and confidence filtering

```
Require: Trained model $\mathcal{M}$,
    labelled dataset $\mathcal{D}_L = \{(V^i, T^i, S^i_{\text{prior}}, S^i_{\text{gt}})\}$,
    audio-aligned dataset $\mathcal{D}_A = \{(V^i, T^i, S^i_{\text{prior}})\}$,
    confidence threshold $\tau$,
    number of self-training epochs $N$
for $n = 1$ to $N$ do
    Initialise pseudo-labelled set $\mathcal{D}_P \leftarrow \emptyset$
    for each $(V, T, S_{\text{prior}}) \in \mathcal{D}_A$ do
        Compute probability $\{S_{\text{pr},t}\}_{t=1}^{T} \leftarrow \mathcal{M}(V, T, S_{\text{prior}})$
        Generate pseudo-label $\hat{S}_{\text{pr}}$ from $S_{\text{pr}}$
        Compute confidence score $c \leftarrow \max_t S_{\text{pr},t}$
        if $c \geq \tau$ then
            $\mathcal{D}_P \leftarrow \mathcal{D}_P \cup \{(V, T, S_{\text{prior}}, \hat{S}_{\text{pr}})\}$
        end if
    end for
    Fine-tune $\mathcal{M}$ on $\mathcal{D}_P$
    Fine-tune $\mathcal{M}$ on $\mathcal{D}_L$
end for
return Final model $\mathcal{M}$
```

Ablation study on selective alignment loss. In Table XII, we analyse the effects of negative alignment loss $\mathcal{L}_{neg}$ and relative alignment loss $\mathcal{L}_{rel}$ that constitute the proposed selective alignment (SA) loss. When using only $\mathcal{L}_{rel}$, we observe a slight degradation in model performance. This decrease is expected because it unnecessarily restricts the model's ability to predict a balanced ratio of 0 or 1 without considering negative samples. On the other hand, when using only $\mathcal{L}_{neg}$, the model shows higher frame-level accuracy and lower F1-scores. This is because it helps the model predict all frames as 0 for negative pairs, thereby predicting the alignment span more tightly. Ultimately, the combination allows the model to benefit from the strengths of each loss component, resulting in improved frame-level accuracy and F1-scores.

Ablation study on self-training. As shown in Table XIII, we conduct another round of self-training, resulting in marginal performance improvements across most metrics. This outcome underscores that even if the self-training process is performed only once, it shows sufficiently saturated performance.

TABLE XIII: **Ablation study on self-training.** The performance improvement from self-training over 2 epochs is marginal. This demonstrates that self-training for just 1 epoch is sufficient to achieve satisfactory performance.

Method	frame-acc (%) ↑	F1@.10 ↑	F1@.25 ↑	F1@.50 ↑
Ours (self-training for 1 epoch)	77.22	81.39	75.03	63.81
Ours (self-training for 2 epoch)	77.39	81.27	75.19	63.96

D. Qualitative results

In this subsection, we qualitatively analyse frame-level model predictions to verify the feasibility of the proposed method. To do this, we visualise the predicted timeline and alignment probability in Fig. 3. We sequentially show the output for the shifted audio subtitle, the SAT model, our model without selective alignment (denoted as *Ours w/o SA loss*), and our full model, including ground truth timing.

Predicted alignment timeline. In the upper part of Fig. 3, while the shifted subtitle prior provides an approximate position, it is largely misaligned with the ground truth. The SAT model exhibits a stronger bias towards the prior alignment, failing to align well with the text query. In contrast, both our model without selective loss and the full model show improved alignment over the SAT model. This improvement stems from a better understanding of the relationship between text queries and video content, particularly in sign language linguistics. Upon closer inspection, applying selective alignment loss further improves alignment with the ground truth. This is attributed to the model's ability to fine-tune alignment by reducing attention to frames unrelated to the text query.

Alignment probability visualisation. In the lower part of Fig. 3, we further visualise the alignment probabilities to analyse where the models focus on. The SAT model tends to emphasise a timeline biased towards the input prior. The model trained without selective alignment loss exhibits better alignment performance than the SAT model but shows low confidence for some of the positive frames. Finally, our full model showcases reduced temporal attention for frames that do not correspond to the text query, while increasing attention for matching segments. This results in predictions that closely align with the ground truth, indicating the effectiveness of the selective alignment loss. This loss facilitates a better understanding of the correlation between sign language and text, particularly in accurately aligning video segments with corresponding textual content.

V. Conclusion

In conclusion, this work presents a novel framework aimed at enhancing the accuracy of sign annotations compared to existing baselines. we address the oversight in previous research regarding the distinct grammatical disparities between sign and spoken languages. Moreover, we tackle the challenges posed by noisy and weak supervision inherent in datasets collected from TV broadcasts. To mitigate noisy supervision, we introduce a selective alignment loss mechanism, penalising misalignments unrelated to text queries and ensuring precise alignment with subtitles. Additionally, we alleviate weak supervision by implementing a self-training strategy, which leverages pseudo-labels generated by our model to further enhance performance. Our findings pave the way for future research in sign language processing and interpretation, ultimately contributing to improved accessibility and inclusivity for the Deaf and hard-of-hearing community.

Limitations. Our experiments were restricted to video inputs of no longer than 20 seconds. Consequently, if the temporal discrepancy between the audio-aligned subtitle labels and the ground truth annotations exceeds this limit, our model inevitably produces inaccurate alignment results. This constraint arises due to memory limitations and the difficulty of modelling long-range temporal dependencies. Nevertheless, we anticipate resolving this limitation in future work by employing techniques such as sliding window inference or chunk-based processing with overlap, which can effectively cover longer sequences without sacrificing local context.

In addition to this, Table V provides an analysis of alignment performance with respect to the characteristics of the input during inference. It shows that the model tends to fail more frequently on samples with relatively short textual content. When such shorter texts are excluded from the evaluation, a noticeable performance improvement is observed. This implies that the length and richness of input language information significantly influence alignment quality. One plausible explanation is that shorter text segments often result in less accurate priors, as subtitle timing becomes less reliable and more sensitive to noise at shorter durations. Furthermore, limited textual input reduces the amount of contextual information available for the model to disambiguate the correct alignment.

These observations highlight a fundamental limitation of relying on subtitle-derived priors during alignment. While the prior helps narrow down the temporal search space, its effectiveness is contingent upon the quality and length of the input text. As a result, the model's robustness is undermined when the prior is poorly aligned or the textual content is insufficient to provide meaningful guidance.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT, RS-2025-02263977, Development of Communication Platform supporting User Anonymization and Finger Spelling-Based Input Interface for Protecting the Privacy of Deaf Individuals). The authors would like to thank Andrew Zisserman, Gül Varol, and Liliane Momeni for their guidance and valuable feedback on the experimental analysis, and also thank Zifan Jiang for providing instructions on the use of the WMT dataset.

References

[1] L. Hu, L. Gao, Z. Liu, and W. Feng, "Continuous sign language recognition with correlation network," in *Proc. CVPR*, 2023.

[2] Y. Chen, F. Wei, X. Sun, Z. Wu, and S. Lin, "A simple multi-modality transfer learning baseline for sign language translation," in *Proc. CVPR*, 2022.

[3] Y. Chen, R. Zuo, F. Wei, Y. Wu, S. Liu, and B. Mak, "Two-stream network for sign language recognition and translation," in *NeurIPS*, 2022.

[4] R. Zuo and B. Mak, "C2slr: Consistency-enhanced continuous sign language recognition," in *Proc. CVPR*, 2022.

[5] J. Kan, K. Hu, M. Hagenbuchner, A. C. Tsoi, M. Bennamoun, and Z. Wang, "Sign language translation with hierarchical spatio-temporal graph neural network," in *Proc. WACV*, 2022.

[6] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Multi-channel transformers for multi-articulatory sign language translation," in *Proc. ECCV*, 2020.

[7] N. C. Camgöz, B. Saunders, G. Rochette, M. Giovanelli, G. Inches, R. Nachtrab-Ribback, and R. Bowden, "Content4all open research sign language translation datasets," in *Proc. FG*, 2021.

[8] S. Albanie, G. Varol, L. Momeni, T. Afouras, H. Bull, H. Chowdhury, N. Fox, B. Woll, R. Cooper, A. McParland, and A. Zisserman, "BOBSL: BBC-Oxford British Sign Language Dataset," *arXiv preprint arXiv:2111.03635*, 2021.

[9] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in *Proc. CVPR*, 2018.

[10] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Computer Vision and Image Understanding*, 2015.

[11] H. Zhou, W. Zhou, W. Qi, J. Pu, and H. Li, "Improving sign language translation with monolingual data by sign back-translation," in *Proc. CVPR*, 2021.

[12] G. Varol, L. Momeni, S. Albanie, T. Afouras, and A. Zisserman, "Read and attend: Temporal localisation in sign language videos," in *Proc. CVPR*, 2021.

[13] S. Albanie, G. Varol, L. Momeni, T. Afouras, J. S. Chung, N. Fox, and A. Zisserman, "BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues," in *Proc. ECCV*, 2020.

[14] L. Momeni, G. Varol, S. Albanie, T. Afouras, and A. Zisserman, "Watch, read and lookup: learning to spot signs from multiple supervisors," in *Proc. ACCV*, 2020.

[15] H. Bull, M. Gouiffès, and A. Braffort, "Automatic segmentation of sign language into subtitle-units," in *Proc. ECCV Workshops*, 2020.

[16] H. Bull, T. Afouras, G. Varol, S. Albanie, L. Momeni, and A. Zisserman, "Aligning subtitles in sign language videos," in *Proc. ICCV*, 2021.

[17] R. Sutton-Spence and B. Woll, *The linguistics of British Sign Language: an introduction*. Cambridge University Press, 1999.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. ACL*, 2018.

[19] R.-H. Liang and M. Ouhyoung, "A real-time continuous gesture recognition system for sign language," in *Proc. FG*, 1998.

[20] R. Yang and S. Sarkar, "Detecting coarticulation in sign language using conditional random fields," in *Proc. ICPR*, 2006.

[21] H.-D. Yang, S. Sclaroff, and S.-W. Lee, "Sign language spotting with a threshold model based on conditional random fields," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2008.

[22] M. S. P. Santemiz, Oya Aran and L. Akarun., "Automatic sign segmentation from continuous signing via multiple sequence alignment," in *Proc. ICCVW*, 2009.

[23] N. P. Eng-Jon Ong, Oscar Koller and R. Bowden, "Sign spotting using hierarchical sequential patterns with temporal intervals," in *Proc. CVPR*, 2014.

[24] P. Buehler, M. Everingham, and A. Zisserman, "Employing signed TV broadcasts for automated learning of British sign language," in *Proc. Workshop on the Representation and Processing of Sign Languages*, 2010.

[25] T. Pfister, J. Charles, and A. Zisserman, "Large-scale Learning of Sign Language by Watching TV (Using Co-occurrences)," in *Proc. BMVC.*, 2013.

[26] D. Li, X. Yu, C. Xu, L. Petersson, and H. Li, "Transferring cross-domain knowledge for video sign language recognition," in *Proc. CVPR*, 2020.

[27] G. Varol, L. Momeni, S. Albanie, T. Afouras, and A. Zisserman, "Scaling up sign spotting through sign language dictionaries," *Intl. Journal of computer vision*, 2022.

[28] O. Koller, O. Zargaran, H. Ney, and R. Bowden, "Deep sign: Hybrid CNN-HMM for continuous sign language recognition," in *Proc. BMVC.*, 2016.

[29] O. Koller, S. Zargaran, and H. Ney, "Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs," in *Proc. CVPR*, 2017.

[30] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006.

[31] Z. Niu and B. Mak, "Stochastic Fine-Grained Labeling of Multi-state Sign Glosses for Continuous Sign Language Recognition," in *Proc. ECCV*, 2020.

[32] R. Cui, H. Liu, and C. Zhang, "A deep neural framework for continuous sign language recognition by iterative training," *IEEE Trans. on Multimedia*, 2019.

[33] Y. Jang, Y. Oh, J. W. Cho, D.-J. Kim, J. S. Chung, and I. S. Kweon, "Signing outside the studio: Benchmarking background robustness for continuous sign language recognition," in *Proc. BMVC.*, 2022.

[34] Y. Min, A. Hao, X. Chai, and X. Chen, "Visual Alignment Constraint for Continuous Sign Language Recognition," in *Proc. ICCV*, 2021.

[35] J. Ahn, Y. Jang, and J. S. Chung, "Slowfast Network for Continuous Sign Language Recognition," in *Proc. ICASSP*, 2024.

[36] N. Adaloglou, T. Chatzis, I. Papastratis, A. Stergioulas, G. T. Papadopoulos, V. Zacharopoulou, G. J. Xydopoulos, K. Atzakas, D. Papazachariou, and P. Daras, "A comprehensive study on deep learning-based methods for sign language recognition," *IEEE Trans. on Multimedia*, 2021.

[37] H. Zhou, W. Zhou, Y. Zhou, and H. Li, "Spatial-temporal multi-cue network for sign language recognition and translation," *IEEE Trans. on Multimedia*, 2021.

[38] O. Koller, C. Camgoz, H. Ney, and R. Bowden, "Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2019.

[39] Y. Jang, Y. Oh, J. W. Cho, M. Kim, D.-J. Kim, I. S. Kweon, and J. S. Chung, "Self-sufficient framework for continuous sign language recognition," in *Proc. ICASSP*, 2023.

[40] H. Hu, J. Pu, W. Zhou, H. Fang, and H. Li, "Prior-aware cross modality augmentation learning for continuous sign language recognition," *IEEE Trans. on Multimedia*, 2023.

[41] S. Ham, K. Park, Y. Jang, Y. Oh, S. Yun, S. Yoon, C. J. Kim, H.-M. Park, and I. S. Kweon, "Ksl-guide: A large-scale korean sign language dataset including interrogative sentences for guiding the deaf and hard-of-hearing," in *Proc. FG*, 2021.

[42] L. Anne Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell, "Localizing moments in video with natural language," in *Proc. ICCV*, 2017.

[43] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "Tall: Temporal activity localization via language query," in *Proc. ICCV*, 2017.

[44] M. Liu, X. Wang, L. Nie, Q. Tian, B. Chen, and T.-S. Chua, "Cross-modal moment localization in videos," in *Proc. ACM MM*, 2018.

[45] S. Zhang, J. Su, and J. Luo, "Exploiting temporal relationships in video moment localization with natural language," in *Proc. ACM MM*, 2019.

[46] J. Chen, X. Chen, L. Ma, Z. Jie, and T.-S. Chua, "Temporally grounding natural sentence in video," in *Proc. EMNLP*, 2018.

[47] Y. Yuan, L. Ma, J. Wang, W. Liu, and W. Zhu, "Semantic conditioned dynamic modulation for temporal sentence grounding in videos," *NeurIPS*, 2019.

[48] D. Zhang, X. Dai, X. Wang, Y.-F. Wang, and L. S. Davis, "Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment," in *Proc. CVPR*, 2019.

[49] M. Soldan, M. Xu, S. Qu, J. Tegner, and B. Ghanem, "Vlg-net: Video-language graph matching network for video grounding," in *Proc. ICCV*, 2021.

[50] S. Zhang, H. Peng, J. Fu, and J. Luo, "Learning 2d temporal adjacent networks for moment localization with natural language," in *Proc. AAAI*, 2020.

[51] Y. Yuan, T. Mei, and W. Zhu, "To find where you talk: Temporal sentence localization in video with attention based location regression," in *Proc. AAAI*, 2019.

[52] S. Ghosh, A. Agarwal, Z. Parekh, and A. Hauptmann, "ExCL: Extractive Clip Localization Using Natural Language Descriptions," in *Proc. ACL*.

[53] H. Zhang, A. Sun, W. Jing, and J. T. Zhou, "Span-based localizing network for natural language video localization," in *Proc. ACL*, 2020.

[54] J. Mun, M. Cho, and B. Han, "Local-global video-text interactions for temporal grounding," in *Proc. CVPR*, 2020.

[55] C. Rodriguez-Opazo, E. Marrese-Taylor, B. Fernando, H. Li, and S. Gould, "DORi: Discovering object relationships for moment localization of a natural language query in a video," in *Proc. WACV*, 2021.

[56] Y.-W. Chen, Y.-H. Tsai, and M.-H. Yang, "End-to-end multi-modal video temporal grounding," in *NeurIPS*, 2021.

[57] J. Woo, H. Ryu, Y. Jang, J. W. Cho, and J. S. Chung, "Let me finish my sentence: Video temporal grounding with holistic text understanding," in *Proc. ACM MM*, 2024.

[58] H. Hu, W. Zhao, W. Zhou, Y. Wang, and H. Li, "Signbert: pre-training of hand-model-aware representation for sign language recognition," in *Proc. CVPR*, 2021.

[59] H. Hu, W. Zhao, W. Zhou, and H. Li, "Signbert+: Hand-model-aware self-supervised pre-training for sign language understanding," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2023.

[60] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023.

[61] M. Bain, J. Huh, T. Han, and A. Zisserman, "Whisperx: Time-accurate speech transcription of long-form audio," in *Proc. Interspeech*, 2023.

[62] F. Landini, J. Profant, M. Diez, and L. Burget, "Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: theory, implementation and analysis on standard tasks," *Computer Speech & Language*, 2022.

[63] E. Z. Xu, Z. Song, S. Tsutsui, C. Feng, M. Ye, and M. Z. Shou, "Ava-avd: Audio-visual speaker diarization in the wild," 2022.

[64] D. Kwak, J. Jung, K. Nam, Y. Jang, J.-W. Jung, S. Watanabe, and J. S. Chung, "Voxmm: Rich transcription of conversations in the wild," in *Proc. ICASSP*, 2024.

[65] B. Korbar, J. Huh, and A. Zisserman, "Look, listen and recognise: Character-aware audio-visual subtitling," in *Proc. ICASSP*, 2024.

[66] J. Huh and A. Zisserman, "Character-aware audio-visual subtitling in context," in *Proc. ACCV*, 2024.

[67] M. Kim, C. W. Kim, and Y. M. Ro, "Deep visual forced alignment: learning to align transcription with talking face video," in *Proc. AAAI*, 2023.

[68] Y. He, L. Yang, and S. Wang, "Enhancing visual forced alignment with local context-aware feature extraction and multi-task learning," in *Proc. ICASSP*, 2025.

[69] A. Farhadi and D. Forsyth, "Aligning ASL for statistical translation using a discriminative word model," in *Proc. CVPR*, 2006.

[70] I. Farag and H. Brock, "Learning motion disfluencies for automatic sign language segmentation," in *Proc. ICASSP*, 2019.

[71] K. Renz, N. Stache, S. Albanie, and G. Varol, "Sign segmentation with temporal convolutional networks," in *Proc. ICASSP*, 2021.

[72] B. Gebrekidan Gebre, P. Wittenburg, and T. Heskes, "Automatic signer diarization-the mover is the signer approach," in *Proc. CVPR Workshops*, 2013.

[73] B. G. Gebre, P. Wittenburg, T. Heskes, and S. Drude, "Motion history images for online speaker/signer diarization," in *Proc. ICASSP*, 2014.

[74] S. Albanie, G. Varol, L. Momeni, T. Afouras, A. Brown, C. Zhang, E. Coto, N. Camgoz, B. Saunders, A. Dutta *et al.*, "Signer diarisation in the wild," *Technical Report*, 2021.

[75] N. Cherniavsky, R. E. Ladner, and E. A. Riskin, "Activity detection in conversational sign language video for mobile telecommunication," in *Proc. FG*, 2008.

[76] M. Borg and K. P. Camilleri, "Sign language detection "in the wild" with recurrent neural networks," in *Proc. ICASSP*, 2019.

[77] A. Moryossef, I. Tsochantaridis, R. Aharoni, S. Ebling, and S. Narayanan, "Real-time sign language detection using human pose estimation," in *Proc. ECCV*, 2020.

[78] F. M. Shipman, S. Duggina, C. D. Monteiro, and R. Gutierrez-Osuna, "Speed-accuracy tradeoffs for detecting sign language content in video sharing sites," in *Proc. ACM SIGACCESS*, 2017.

[79] D. Li, C. Xu, X. Yu, K. Zhang, B. Swift, H. Suominen, and H. Li, "Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation," *NeurIPS*, 2020.

[80] Y. Jang, H. Raajesh, L. Momeni, G. Varol, and A. Zisserman, "Lost in translation, found in context: Sign language translation with contextual cues," in *Proc. CVPR*, 2025.

[81] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, 2017.

[82] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *NeurIPS*, 2020.

[83] M. Müller, S. Ebling, E. Avramidis, A. Battisti, M. Berger, R. Bowden, A. Braffort, N. C. Camgöz, C. España-Bonet, R. Grundkiewicz *et al.*, "Findings of the first wmt shared task on sign language translation (wmt-slt22)," in *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 2022.

[84] C. Raude, K. R. Prajwal, L. Momeni, H. Bull, S. Albanie, A. Zisserman, and G. Varol, "A tale of two languages: Large-vocabulary continuous sign language recognition from spoken language supervision," *arXiv*, 2024.

[85] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

[86] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proc. CVPR*, 2017.

[87] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2014.

[88] L. Momeni, H. Bull, K. Prajwal, S. Albanie, G. Varol, and A. Zisserman, "Automatic dense annotation of large-vocabulary sign language videos," in *Proc. ECCV*, 2022.

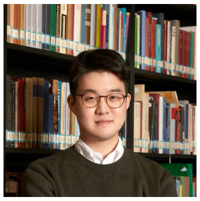
[89] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, 2020.

[90] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proc. CVPR*, 2022.

Youngjoon Jang received the Ph.D. degree in electrical engineering from Korea Advanced Institute Science and Technology, Daejeon, South Korea. He is currently a Postdoctoral Research Fellow at Korea Advanced Institute Science and Technology. His research interests include automatic sign language recognition/translation, video understanding, video generation, and multi-modal analysis.

Jeongsoo Choi received the B.S. degree in electrical engineering from Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2020. He is currently pursuing the Ph.D. degree in electrical engineering at Korea Advanced Institute of Science and Technology. His research interests include deep learning, image/video analysis, speech synthesis, and multi-modal analysis.

Junseok Ahn received the B.S. degree in electrical engineering from Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2023. He is currently pursuing the Ph.D. degree in electrical engineering at Korea Advanced Institute of Science and Technology. His research interests include deep learning, audio/video generation, sign language recognition/translation, and multi-modal analysis.



Joon Son Chung is an associate professor at Korea Advanced Institute of Science and Technology, where he is directing research in speech processing, computer vision and machine learning. He received the D.Phil. in Engineering Science from the University of Oxford.