Aligning Subtitles in Sign Language Videos

Hannah Bull[1*] Triantafyllos Afouras[2*] Gül Varol[2,3]
Samuel Albanie[2,4] Liliane Momeni[2] Andrew Zisserman[2]
[1] LISN, Univ Paris-Saclay, CNRS, France
[2] Visual Geometry Group, University of Oxford, UK
[3] LIGM, École des Ponts, Univ Gustave Eiffel, CNRS, France
[4] Department of Engineering, University of Cambridge, UK
hannah.bull@lisn.upsaclay.fr; {afourast,gul,albanie,liliane,az}@robots.ox.ac.uk
https://www.robots.ox.ac.uk/~vgg/research/bslalign/

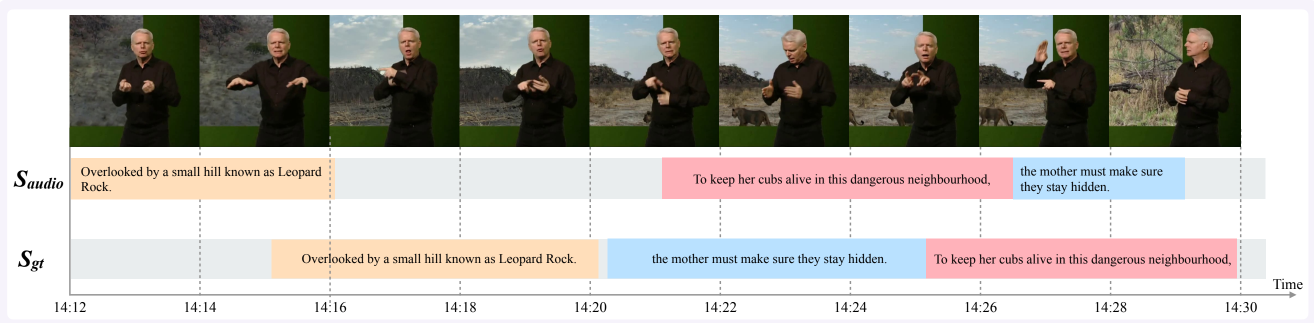


Figure 1: **Subtitle alignment**: We study the task of aligning subtitles to continuous signing in sign language interpreted TV broadcast data. The subtitles in such settings usually correspond to and are aligned with the audio content (top: audio subtitles, S_{audio}) but are unaligned with the accompanying signing (bottom: Ground Truth annotation of the signing corresponding to the subtitle, S_{gt}). This is a *very challenging* task as (i) the *order* of subtitles varies between spoken and sign languages, (ii) the *duration* of a subtitle differs considerably between signing and speech, and (iii) the signing corresponds to a *translation* of the speech as opposed to a transcription.

Abstract

The goal of this work is to temporally align asynchronous subtitles in sign language videos. In particular, we focus on sign-language interpreted TV broadcast data comprising (i) a video of continuous signing, and (ii) subtitles corresponding to the audio content. Previous work exploiting such weakly-aligned data only considered finding keyword-sign correspondences, whereas we aim to localise a complete subtitle text in continuous signing. We propose a Transformer architecture tailored for this task, which we train on manually annotated alignments covering over 15K subtitles that span 17.7 hours of video. We use BERT subtitle embeddings and CNN video representations learned for sign recognition to encode the two signals, which interact through a series of attention layers. Our model outputs frame-level predictions, i.e., for each video frame, whether it belongs to the queried subtitle or not. Through extensive evaluations, we show substantial improvements over existing alignment baselines that do not make use of subtitle text embeddings for learning. Our automatic alignment model opens up possibilities for advancing machine translation of sign languages via providing continuously synchronized video-text data.

1. Introduction

Sign languages constitute a key form of communication for Deaf communities [50]. Our goal in this paper is to temporally localise subtitles in continuous signing video. Automatic alignment of subtitle text to signing content has great potential for a wide range of applications including assistive tools for education and translation, indexing of sign language video corpora, efficient subtitling technology for signing vloggers[1], and automatic construction of large-scale sign language datasets that support computer vision and linguistic research.

Despite recent advances in computer vision, machine

[*]Equal contribution

[1]Unlike spoken vlogs that benefit from automatic closed captioning on sites such as YouTube, signing vlog creators who wish to provide written subtitles must both translate *and* align their subtitles manually.

translation between continuous signing and written language remains largely unsolved [6]. Recent works [11, 12] have shown promising translation results, but to date these have been achieved only in *constrained* settings where continuous signing is *manually pre-segmented* into clips, with each clip associated to a written sentence from a *limited vocabulary*. Two key bottlenecks for scaling up translation to continuous signing depicting unconstrained vocabularies are (i) the segmentation of signing into sentence-like units, and (ii) the availability of large-scale sign language training data.

Manual alignment of subtitles to sign language video is tedious – an expert fluent in sign language takes approximately 10-15 hours to align subtitles to 1 hour of continuous sign language video. In this work, we focus on the task of aligning a particular known subtitle within a given temporal signing window. We explore this task in the context of sign language interpreted TV broadcast footage – a readily available and large-scale source of data – where the subtitles are synchronised with the audio, but the corresponding sign language translations are largely unaligned due to differences between spoken and sign languages as well as lags from the live interpretation.

Subtitle alignment to continuous signing remains a *very challenging* task. First, sign languages have grammatical structures that vary considerably from those of spoken languages [50], and as a result the *ordering* of words within a subtitle as well as the subtitles themselves is often not maintained in the signing (see Fig. 1). Second, the *duration* of a subtitle varies considerably between signing and speech due to differences in speed and grammar. Third, the signing corresponds to a *translation* of the speech that appears in the subtitles as opposed to a transcription: there is no direct one-to-one mapping between subtitle words and signs produced by interpreters, and entire subtitles may not be signed.

Previous work exploiting such weakly-aligned data has mainly focused on finding sparse correspondences between keywords in the subtitle and individual signs [3, 39, 53], as opposed to localising the start and end times of a complete subtitle text in continuous signing. Though, as we show, localising isolated signs identified by keyword spotting nevertheless forms a useful pretraining task for full subtitle alignment. Most closely related to our work, Bull et al. [9] consider the task of segmenting a continuous signing video into subtitle units purely based on body keypoints. In fact, similarly to speech which can be segmented based on prosodic cues such as pauses, sign sentence boundaries can *to an extent* be detected through visual cues such as lowering the hands, head movement, pauses, and facial expressions [24]. However, as shown in our evaluations in Sec. 4, such approaches based on prosody-only perform poorly in our setting, where subtitles do not necessarily correspond to complete sign sentences with clear visual boundaries.

In this paper, we instead propose to use *the subtitle text as an additional signal* for better alignment. We make the following three contributions: (1) we show that encoding the subtitle text as input to the alignment model significantly improves the temporal localisation quality as opposed to only relying on visual cues to segment continuous sign language videos into subtitle units; (2) we design a novel formulation for the subtitle alignment task based on Transformers; and (3) we present a comprehensive study ablating our design choices and provide promising results for this new task when evaluating on unseen signers and content.

2. Related Work

For a recent comprehensive survey about sign language recognition and translation, see [32]. Here, we review relevant works on temporal localisation at the levels of individual signs and sequences, in addition to more general temporal alignment methods from the literature.

Temporal localisation of individual signs. A rich body of work has considered the task of localising sparse sign instances in continuous signing, often referred to as "sign spotting". Early efforts using signing gloves [37] were followed by methods employing hand-crafted visual features to represent the hands, face and motion that were integrated with CRFs [58, 59], HMMs [46] and HSP Trees [42]. Several studies have sought to employ subtitles as weak supervision for learning to localise and classify signs, using apriori mining [18] and multiple-instance learning [7, 8, 43]. More recent work has leveraged cues such as mouthings [3] and visual dictionaries [39] and by making use of deep neural network features with sliding window classifiers [36] and attention learned via a proxy translation task [53]. In deviation from these works, our objective is to localise complete subtitle units, rather than individual signs.

Temporal localisation of sign sequences. The alignment of subtitles to continuous signing was considered in creative early work by combining cues from multiple sparse correspondences [23], but under the assumption that ordering of words in subtitles are preserved in the signing (which does not hold in our problem setting). Other sequence-level sign language temporal localisation tasks that have received attention in the literature include category-agnostic sign segmentation [22, 44], active signer detection [5, 17, 40, 49] and diarisation [2, 26, 27]—each considers a temporal granularity that differs from subtitle units. Most closely related to our work, Bull et al. [9] employ a keypoint-based model to segment continuous signing into sentence-like units without knowledge of the written subtitles during inference. Our approach relaxes this assumption and considers instead the practical scenario in which we assume access to the written subtitle to be aligned. We compare our approach with theirs in Sec. 4.

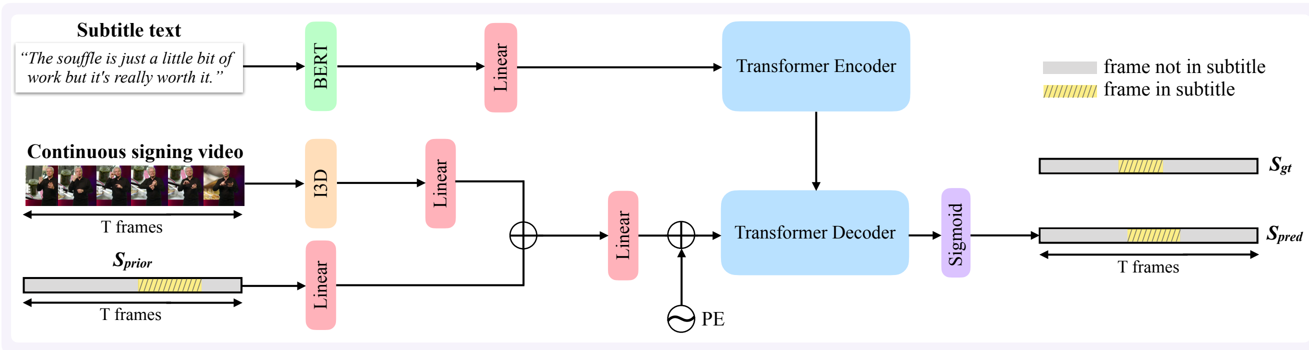


Figure 2: **SAT model overview:** We input to our model (i) token embeddings of the subtitle text we wish to align, (ii) a sequence of video features extracted from a continuous sign language video segment and (iii) the shifted temporal boundaries of the audio-aligned subtitle, S_{prior}. Using these inputs, the model outputs a vector of values between 0 and 1 of length T. Its first and last values above a threshold τ delimit the predicted temporal boundaries for the query subtitle. The location of the subtitle with respect to the window is represented in dashed yellow.

Continuous sign language recognition. Hybrid models coupling CNNs with HMMs [33, 34], attention mechanisms [31] and CTC losses [10, 16] have been studied for continuous sign language recognition, with recent extensions to sequence-to-sequence models [11] and Transformers [12, 35] to tackle the task of sign language translation. These models produce either implicit or explicit alignments over a signing sequence corresponding to a sentence. However, these approaches have only been demonstrated to work on *pre-segmented* sentences of signing [11].

Aligning bodies of text to video. The Dynamic Time Warping (DTW) algorithm [41] has been applied to the problem of aligning sequences of movies to transcripts [21, 45] and plots synopses [51] using cues such as character recognition and subtitle content. It has also been successfully applied to the problem of aligning generic text descriptions against untrimmed video [4]. While effective, these methods require the preservation of sequence ordering across modalities, which does not hold in our problem setting. We nevertheless show in Sec. 3 how DTW can be used as a secondary stage of processing that resolves conflicting local alignments on the re-ordered subtitle prediction timings via a global objective. The fixed ordering assumption is relaxed by the work of [52], which aligns book chapters to video scenes. Their approach, however, which works through matching sparse character identifications against specific shots, is not applicable in our setting where shot boundaries do not provide a natural segmentation of the signing content.

Natural language grounding in videos. Our work is also related to the task of natural language grounding, which aims to locate a temporal segment within an untrimmed video sequence corresponding to a given natural language query. Existing methods have considered two-stage *propose and rank* approaches [25, 30, 38, 56], iterative grounding agents trained with reinforcement learning [29, 55] and single-stage regression models [15, 28, 60, 61]. Our proposed subtitle alignment task differs from natural language grounding in three ways: (i) The signing content is more *fine-grained*—the visual appearance of a signing sequence remains very similar across frames, necessitating nuanced recognition of body dynamics; (ii) Differently from language grounding, each subtitle to be aligned comes with its own reference location, providing an instance-specific prior over the start time and duration. As we show in Sec. 4, our effective use of this reference is important to achieving good performance, and our model is specifically designed to take advantage of this cue; (iii) Subtitles occupy mutually exclusive temporal regions, a property that we further exploit to improve alignment quality, but that does not hold in general for natural language grounding.

3. Method

In this section, we describe our Transformer-based subtitle alignment model operating on a single subtitle and a short video segment (Sec. 3.1), our pretraining on sparse sign spottings (Sec. 3.2), and our final step that globally adjusts multiple subtitles in a long video using DTW (Sec. 3.3).

Problem formulation. As inputs to the model, we provide (i) token embeddings of the subtitle text we wish to align to signing, (ii) a sequence of video features extracted from a continuous sign language video segment, as well as (iii) prior estimates of the temporal boundaries for the given query, which we refer to as S_{prior}. The latter is provided as an approximate location and duration cue of the signing-aligned subtitle. Using these inputs, we predict a binary vector of the same length as the video features, where a consecutive sequence of 1s denotes the temporal location of the subtitle.

3.1. Subtitle Aligner Transformer

The core of our model is a Transformer [54], as shown in Fig. 2, which we refer to as Subtitle Aligner Transformer (SAT). In contrast to the common approach of feeding video frames as input to the encoder [13, 19], we input the video frames to the *decoder* side in order for the model to learn the association between the frame-level features and the output vector of the same duration. We first describe the structure of the Transformer, and then the text and video feature extraction. Additional implementation details are provided in App. Sec. B.

Encoder. The input to the encoder is a sequence of text embeddings corresponding to the subtitle we wish to align. Positional encodings are not used on the encoder side of the Transformer since the text embeddings (see below) already contain positional information. The encoder is a stack of Transformer layers, each containing a multi-head attention mechanism followed by a feedforward network and embedding dimensionalities of size d_{model}.

Decoder. The decoder is a stack of Transformer layers that attend on the encoded sequence.[2] The input to the decoder consists of a sequence of video features encoding the visual signing information from the video, as well as a binary vector representing a prior estimate of the location of the signing-aligned subtitle (S_{prior}). Positional encodings are added to the decoder input in order for the model to exploit the temporal ordering of the signing. The final layer of the model is a linear layer with a sigmoid activation which outputs T predictions in the range $[0, 1]$ one for each video frame. Values of this output vector, S_{pred}, that are above a threshold τ correspond to the predicted temporal location of the queried subtitle text.

Text features. Each subtitle is encoded using a BERT [20] model, pretrained on a large text corpus with a masked language modelling task, to produce a sequence of 768-dimensional vectors, one for each token in the sentence. To match the input dimension of the encoder Transformer, these embeddings are first linearly projected to d_{model}.

Video features. The visual features are 1024-dimensional embeddings extracted from the I3D [14] sign classification model made publicly available by the authors of [53]. The features are pre-extracted over sign language video segments. A visual feature sequence of length T is used as input to the model.

Prior position encoding. Besides the video features, the input to the decoder also includes a subtitle timing estimate as a prior position and duration cue. This prior estimate is encoded as a binary vector of length T, where 1 indicates that the associated video frame is within the temporal boundaries of the subtitle, and 0 otherwise. The video and prior inputs are fused via concatenation before being passed as input to the decoder. Before the concatenation both inputs are linearly projected to the same dimension. The fusion output is finally projected to d_{model} in order to be input to the Transformer decoder.

Training objective. The model is trained with a binary cross entropy loss between the predicted vector and the ground truth S_{gt} of the signing-aligned subtitle within the video segment:

$$\mathcal{L} = -\frac{1}{T}\sum_{t=1}^{T} S_{gt}^{t} \log S_{pred}^{t} + (1 - S_{gt}^{t}) \log(1 - S_{pred}^{t}).$$

3.2. Word pretraining with individual sign locations

SAT is designed for alignment of subtitles to video signing streams. However, the same architecture can be used without any alterations to align smaller text units, e.g. single words. Given that we have access to sparse sign annotations from mouthings [3] and dictionary exemplars [39], we can use these to initialise the model weights and incorporate this knowledge via a potentially easier single-sign spotting task. We obtain timings of the sparse word-level annotations and assume a fixed single-second width as the precise sign boundaries are not available. The model is then trained to spot the single sign occurrence within a video window of size T. In our experiments, we demonstrate the advantages of such a pretraining strategy.

3.3. Global alignment with DTW

Our model does not take into account global information from the length of the video (e.g. 1-hour), rather it looks for signing associated to a given subtitle within a short temporal window T (e.g. 20-seconds). Hence, there may be overlaps between predictions for different subtitles; we resolve these overlap conflicts using DTW [41]. We find an order-preserving global alignment from all elements of a sequence of video frames to all elements of sequence of subtitles, maximising the sum of sigmoid outputs of our model in our cost function for each subtitle query.

As DTW aligns all frames in a video sequence to subtitles, we select all frames of the signing video which are likely to be associated with subtitle queries. Specifically, we select all frames associated to an output score over τ_{dtw}. In the case where our model outputs only values below τ_{dtw} for a particular subtitle, we instead select all frames within the prior location S_{prior}.

We order the subtitles by the mid-point of their predicted temporal location. This allows the predicted subtitles to follow a different order to the original subtitles, because the order of phrases in the sign language interpretation does not necessarily follow the order of phrases of the written English subtitles (see App. Sec. C for further details).

We construct a cost matrix of dimension (i) the number of frames by (ii) the number of subtitles, and with entries of $1 - p_{ij}$, where p_{ij} is the sigmoid output corresponding to frame i with subtitle j as the encoder input. We apply the

[2]Note: There is no auto-regression.

DTW algorithm to this cost matrix of aligning video frames to subtitles. This maximises the overall sum of the sigmoid outputs of the model under the ordering and allocation constraints of DTW.

If not otherwise mentioned, our full SAT model uses DTW postprocessing.

4. Experiments

In this section, we first give implementation details (Sec. 4.1) and describe the datasets and evaluation metrics used in this work (Sec. 4.2). We then compare the results of the proposed SAT model against strong baselines (Sec. 4.3) and present a series of ablation studies (Sec. 4.4). Next, we demonstrate the performance of our model on additional datasets (Sec. 4.5). Finally, we provide qualitative results and discuss limitations (Sec. 4.6).

4.1. Implementation details

Architecture. For both the encoder and the decoder we use 2 identical Transformer layers with 2 heads and size $d_{model} = 512$ each.

Backbone pretraining. The I3D model is pretrained to perform 1064-way classification across the sign spotting instances with mouthings [3] and dictionary exemplars [39] (further details can be found in [53]). The model is then frozen and used to densely pre-extract visual features with stride 1 over the clips of the datasets.

Prior input selection. As the prior estimate input S_{prior} we use the temporal location of the audio-aligned subtitle S_{audio} shifted by +3.2 seconds. This value, which we denote with S^{+}_{audio}, corresponds to the average temporal shift between the audio-aligned subtitles S_{audio} and the ground truth subtitles S_{gt} in our training data (see Fig. 3a).

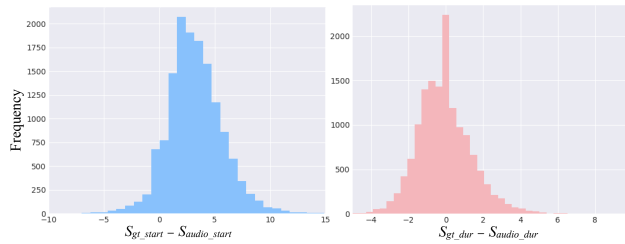
Search windows. During training, we randomly select a search window of 20 seconds around the location of the ground truth subtitle S_{gt}, select the densely extracted video features for this window, and temporally subsample them by a factor of 4. All videos are sampled at 25 FPS, therefore this results in $T = 125$ frames. During testing, we select a search window of the same length centered around the shifted subtitle location S^{+}_{audio}. An ablation study on the window size can be found in App. Sec. D.

Text augmentation. During training, we augment the text query inputs randomly to reduce overfitting. For 50% of the samples, we shuffle the word order and add or delete up to two words.

Hyper-parameters. We set thresholds τ to 0.5, τ_{dtw} to 0.4. Further details are provided in App. Sec. B.

4.2. Data and evaluation metrics

Statistics on the number of videos, hours, subtitles and vocabulary of each of our training and evaluation datasets are provided in Tab. 1. We briefly describe each dataset and provide further details in App. Sec. A.



(a) (b)

Figure 3: **S_{gt} vs. S_{audio}**: We plot the distribution of temporal shifts between ground-truth (S_{gt}) and audio-aligned (S_{audio}) subtitles on the training split of the BSL-1K$_{aligned}$ dataset by showing the differences in subtitle (a) start times and (b) duration. We observe the difficulty of the subtitle alignment task: (i) there is no fixed shift between ground-truth and audio-aligned subtitle timings, and (ii) the subtitle duration varies between spoken and signed languages.

		#vids.	#hours	#subs	#inst.	Vocab.	OOV
BSL-1K$_{aligned}$	Train	20	14.4	13.8K	128.1K	8.6K	\
	Test	4	3.3	2.0K	18.6K	2.8K	0.7K
BSL Corpus	Train	191	22.9	33.7K	261.5K	7.5K	\
	Val	15	1.5	2.6K	18.1K	1.8K	0.2K
	Test	21	2.6	3.8K	27.3K	2.4K	0.4K
BOBSL	Test	36	30.1	28.5K	248.9K	14.3K	8.9K

Table 1: **Datasets:** number of videos, hours, subtitles, word instances, vocabulary size and number of out-of-vocabulary (OOV) words.

BSL-1K$_{aligned}$ is a subset of BSL-1K [3], covering 24 different television programmes (food, nature, travel and lifestyle documentaries). The subtitles were originally aligned to the audio, but we have manually aligned them to the signing. The unaligned subtitles (i.e. those that are synchronised with the audio track, rather than the signing) differ from the signing-aligned subtitles in both start time and duration. In particular, Fig. 3, shows that there is no fixed shift or temporal scaling that can be consistently applied to transform audio-synchronised subtitles to their signing-aligned counterparts. We note that the differences exhibit an approximately Gaussian distribution, with the exception of an accentuated peak at 0 in Fig. 3b; we attribute this to the fact that if the duration of the subtitle is approximately correct, annotators tend to not further refine the boundaries.

BSL Corpus [47, 48] is a public dataset of videos of deaf signers gathered from several regions across the UK and accompanied by a variety of linguistic annotations. Unlike BSL-1K, the subtitles in this dataset are aligned to signing, and the translation direction is from sign language to English. We therefore simulate unaligned data by perturbing the subtitle locations in our experiments.

BOBSL is a dataset similar to BSL-1K$_{aligned}$ in style and content. For a subset of the test set of BOBSL, we manually align the original audio-aligned subtitles to the signing. We release these annotations for research purposes. More details are provided in App. Sec. A.

Evaluation metrics. We consider two main evaluation metrics: (i) frame-level accuracy, and (ii) $F1$-score. For the $F1$-score, hits and misses of subtitle alignment to sign language video are counted under three temporal overlap thresholds (IoU $\in \{0.1, 0.25, 0.50\}$) between predicted S$_{pred}$ and manually aligned S$_{gt}$ subtitles, denoted as F1@.10, F1@.25, F1@.50, respectively.

4.3. Comparison to baselines

Simple temporal shift baseline (S$^{+}_{audio}$). As a first baseline we use the shifted audio-aligned subtitles S$^{+}_{audio}$. Only a third of the shifted-audio subtitles S$^{+}_{audio}$ have more than 50% overlap (IoU) with the ground truth aligned subtitles.

Prosodic cues baseline (Bull et al. [9]). We compare to the state of the art on subtitle-unit segmentation, which is a model based on 2D body keypoints. In contrast to our framework, this method only uses visual prosodic cues and does not use semantic information from the query subtitle. It has been trained on a large-scale sign language corpus with aligned subtitles, and the pretrained model is public. The model consists of ST-GCN [57] and BiLSTM layers and segments sign language video into subtitle units. However, this is a different task than alignment, i.e. segments have no correspondence to subtitles. To obtain an association from each predicted segment to a subtitle, we align the shifted subtitles S$^{+}_{audio}$ to a subtitle-unit segmentation of [9] using DTW, where the cost of alignment is the temporal distance.

Heuristic baseline based on sparse sign spottings. Inspired by previous works that approached the alignment task through sparse correspondences [23], we implement a heuristic approach to align the subtitles using a combination of sign spotting and active signer detection. Sign spotting, performed by [3, 39], searches in the temporal vicinity of each audio-synchronised subtitle (the search window is constructed by padding the original subtitle by four seconds at each end) for individual sign instances corresponding to words that appear in the subtitle. From these sparse sign localisations, we perform subtitle alignment in four stages. First, we segment the episode into sequences that contain active signing, following [2]. Second, for any subtitle containing words that were spotted in the signing (assigned a posterior probability of 0.8 or greater by the model of [39]), we shift the subtitle such that its centre falls on the mean position of the spotted signs. Third, we transform all subtitles without spottings by affine transformations such that they fall within the "gaps" between those subtitles that contained spotted signs, while preserving ordering (we use one such transformation per gap). Finally, we expand the duration of subtitles locally (applying a single scaling factor to each subtitle) in left to right ordering, such that they maximally fill the active signing segments predicted by the first stage. We note that only 15% of the subtitles in our test set can be confidently associated to a sign spotting, therefore relying only on sign localisation is expected to be insufficient for subtitle alignment.

Method	frame-acc	F1@.10	F1@.25	F1@.50
S$_{audio}$	44.67	45.82	30.51	12.57
S$^{+}_{audio}$	60.76	71.69	60.74	36.10
Sign-spotting heuristics	61.71	69.23	59.60	36.04
Bull et al. [9]	62.14	73.93	64.25	38.16
SAT (random subtitle)	65.52	70.30	60.36	40.04
SAT w/out DTW	65.81	74.32	64.69	41.27
SAT	**68.72**	**77.80**	**69.29**	**48.15**

Table 2: **Comparison to baselines:** We show significant improvements by training a Subtitle Aligner Transformer (SAT) over several baselines. Moreover, providing a random subtitle as the text input results in poor performance, demonstrating that our model does indeed rely on token embeddings, and does not simply learn prosodic cues to align the subtitles. We obtain a further boost by correcting the overlaps of our predicted subtitles using DTW.

A comparison of our model to the above baselines is given in Tab. 2. The simple temporal shift baseline and the heuristic baseline based on sparse sign spottings perform similarly, but are a significant improvement over the non-shifted subtitles S$_{audio}$. Using prosodic cues through the model of [9] results in a slight improvement over these two baselines. Our model significantly outperforms all baselines by exploiting the subtitle text to find the associated video segment. Indeed, when providing random subtitle text during training, our model is forced to rely on prosodic cues and fails to outperform the baseline F1 scores. Using DTW to resolve overlaps in predicted subtitles boosts our model performance.

4.4. Ablation study

We ablate the effects of inputting the prior estimate S$_{prior}$ $= S^{+}_{audio}$ to the model, modifying the text input to the encoder, pretraining on sign localisation, and alternative model formulations. Some additional ablations are presented in App. Sec. D.

Knowledge of S$_{prior}$. We experiment with several versions of inputs as additional information to the alignment task. Tab. 3 summarises the results. We first observe a significant drop in performance when S$_{prior}$ is not provided (48.15 vs 30.66 F1@.50), suggesting that the position and duration of the corresponding audio content allows an approximate localisation cue, enabling the model to refine this via a series of attention layers. Inputting the 3.2 seconds shifted subtitle timings (S$_{prior}$ $= S^{+}_{audio}$) performs better than inputting the audio-aligned subtitle timings (S$_{prior}$ $= S_{audio}$). Nev-

Additional input	frame-acc	F1@.10	F1@.25	F1@.50
w/out S_{audio}	61.37	59.03	49.35	30.66
w/ S_{audio}	67.81	74.69	66.53	45.10
w/ S^{+}_{audio} 3.2-sec shift	**68.72**	**77.80**	**69.29**	**48.15**
w/ S_{audio} centre position	61.40	58.07	51.13	35.01
w/ S^{+}_{audio} rand. duration	68.61	75.10	66.84	46.72

Table 3: **Inputting S_{prior} variants:** Without information on the approximate position and duration of the subtitle, our model fails to improve upon our baseline methods. In particular, when setting the input S_{prior} to be systematically in the centre of the search window and with the duration of S_{audio}, model performance is poor. When using S^{+}_{audio} in its correct location in the search window, but varying the duration randomly of up to 2s, performance is relatively high. This suggests that position is a stronger cue than duration.

Method	frame-acc	F1@.10	F1@.25	F1@.50
w/o augmentations	67.35	75.72	66.85	45.31
w/ augmentations	**68.72**	**77.80**	**69.29**	**48.15**
w/ aug. + positional enc.	68.21	74.89	67.14	46.36
w/ aug. sentence emb.	66.18	72.99	63.71	41.71

Table 4: **Text ablations:** Our best model uses word embeddings without positional encodings as well as text augmentations during training (shuffling words in 50% of the subtitles, adding and deleting up to 2 words).

ertheless, our model still performs well when the average subtitle lag is unknown and the audio-aligned subtitle timings are used. Moreover, we carry out two additional experiments to investigate whether this cue is more important for providing a position prior or a duration prior. First, we always input the subtitle timing centred with respect to the search window. The poor performance of this model suggests the importance of the position. Second, we preserve the shifted location, but randomly change the input subtitle duration at training time by up to 2s. This slightly reduces the performance, therefore we infer that duration cues are less essential for the model than location cues.

Effect of text input to the encoder. We perform a series of ablations regarding the text encoding, including: no text augmentations, adding extra positional encodings to the BERT text features (as described in App. Sec. B.), and using the sentence embedding only (the output embedding corresponding to the BERT "CLS" token) instead of the sequence of individual token embeddings. Tab. 4 presents the results on BSL-1K$_{aligned}$ with these text ablations. Augmenting the subtitle text improves performance, while adding extra positional encodings or using the sentence embedding degrades performance.

Effect of sign localisation pretraining. As explained in Sec. 3.2, we initially pretrain our model for temporal localisation of individual signs on a large set of word-video training pairs. In Tab. 5, we measure the effect of this pretraining and conclude that it provides a good initialisation for finetuning on long subtitles.

Pretraining	frame-acc	F1@.10	F1@.25	F1@.50
w/o word pretraining	67.26	76.18	66.19	42.47
w/ word pretraining	**68.72**	**77.80**	**69.29**	**48.15**

Table 5: **Pretraining for sign localisation:** By pretraining our model to locate individual words within a given temporal window, we boost performance of subtitle alignment.

Prior input	Loss	frame-acc	F1@.10	F1@.25	F1@.50
shift/scale	shift/scale regress.	59.23	70.55	59.00	33.71
start/end	start/end regress.	60.04	72.20	60.41	34.33
start/end	binary classif.	60.48	74.05	62.75	35.07
binary	binary classif. (SAT)	**68.72**	**77.80**	**69.29**	**48.15**

Table 6: **Model formulation:** We present an ablation where we experiment with a DETR-style Transformer model [13]. Video features are inputs to the Transformer encoder, and the subtitle query is fed to the Transformer decoder. Moreover, on the decoder side, we input either the start and end times or the shift and scale of the shifted subtitles S^{+}_{audio} relative to the temporal window, and use a regression model to predict the true values. This model fails to produce satisfactory results. Changing the regression model to a classification one by instead predicting a binary vector of length T (as in the SAT model) results in a small improvement; however SAT outperforms all the alternative models with a large margin.

Model formulation. We consider an alternative version of the Transformer model, inspired by the DETR model in [13] for object detection in images. This model inputs image features into the Transformer encoder and text query into the Transformer decoder. Similarly, we input the sign language video features into the Transformer encoder. On the decoder side, we input the subtitle text features as well as either (i) the start and end times or (ii) the shift and scale of the shifted subtitles S^{+}_{audio} relative to the temporal window. We then consider the problem of subtitle alignment as a regression problem, and aim to predict (i) the start and end times or (ii) the shift and scale of the subtitle relative to the temporal window. As a further ablation, we also consider the same model architecture (with subtitle features and the start and end times as decoder input), but outputting a fixed binary vector of length T, which we train with a binary classification objective (as in SAT).

The results in Tab. 6 suggest that our proposed approach with video features as input to the Transformer decoder enables significantly better learning, perhaps by providing a one-to-one mapping between video inputs and the frame-wise outputs. Another possible explanation for our proposed model's superiority is that it outputs alignment scores between subtitles and individual frames which allows for better conflict resolution strategies for overlapping subtitle predictions.

4.5. Performance on different datasets

We demonstrate our model's performance on two more datasets: the BSL Corpus [47, 48] and BOBSL [1].

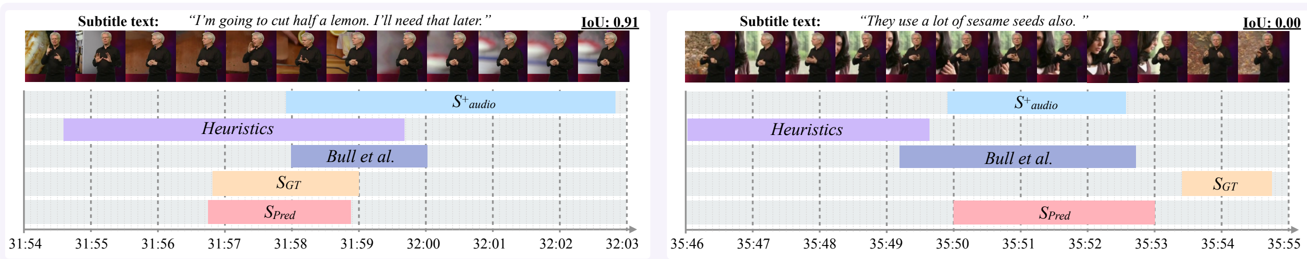


Figure 4: **Qualitative results:** This figure shows short time windows of 9s with shifted audio-aligned subtitles (S^{+}_{audio}), heuristic and Bull et al. [9] baselines, ground truth signing-aligned subtitles (S_{gt}) and our predicted signing-aligned subtitles (S_{pred}). Note that in practice, we input 20 seconds of video during training and testing as our search window. We depict shorter, "zoomed in" 9 second windows here for clearer visualisation. The right shows a failure case.

Rand. perturb. (σ_{pos}, σ_{dur})	Method	frame-acc	F1@.10	F1@.25	F1@.50
(3.5s, 1.5s)	Rand. shift & scale	63.24	37.13	26.54	12.47
	SAT w/out pretrain.	73.73	51.51	43.33	27.98
	SAT pretrain.	75.77	55.55	47.45	32.57
	SAT w/ word pretrain.	**76.29**	**57.65**	**50.35**	**34.54**
(4.5s, 1.5s)	Rand. shift & scale	60.18	29.52	20.61	10.00
	SAT pretrain.	73.69	48.41	41.34	28.06
	SAT w/ word pretrain.	**74.29**	**51.33**	**44.37**	**30.13**
(3.5s, 2s)	Rand. shift & scale	62.62	37.47	26.82	11.87
	SAT pretrain.	75.79	55.31	47.24	32.89
	SAT w/ word pretrain.	**76.00**	**57.86**	**50.43**	**33.79**

Table 7: **BSL Corpus:** We randomly shift and scale the correctly aligned subtitles in BSL Corpus to simulate unaligned data and then use our SAT model to recover the original correct alignments.

BSL Corpus. The subtitles in this dataset are aligned to the sign language, and so we randomly shift and scale the subtitles in order to create artificial training data. We then train our SAT model to learn the correct alignment of subtitles to video in the BSL Corpus. We train the model (i) without any pretraining, (ii) with only word pretraining (on BSL-1K) and (iii) with SAT pretraining on BSL-1K$_{aligned}$. We report results in Tab. 7.

At each subtitle, we apply a random shift following a normal distribution with standard deviation σ_{pos} and a random change of duration of the subtitle also following a normal distribution with standard deviation σ_{dur}. Tab. 7 shows that our model is able to partially recover the correct original alignment. Larger shifts make it more difficult for our model to recover the correct original alignment, but random changes in subtitle duration seems to have less effect. This is consistent with the results in Tab. 3, where changing the duration of S^{+}_{audio} does not greatly impact results. Word pretraining on BSL-1K helps the model, but SAT pretraining on BSL-1K$_{aligned}$ does not. Word pretraining may help the SAT model recognise certain signs in BSL, but domain difference between BSL Corpus and BSL-1K$_{aligned}$ subtitles may explain why SAT pretraining on BSL-1K$_{aligned}$ does not lead to any significant gains on BSL Corpus.

BOBSL. The BOBSL test set allows us to evaluate our model on a larger and more diverse set of videos than the BSL-1K$_{aligned}$ test set. We report results in Tab. 8 and show further qualitative analysis in App. Sec. C.

Method	frame-acc	F1@.10	F1@.25	F1@.50
S_{audio}	23.93	32.94	20.23	7.39
S^{+}_{audio}	50.05	65.48	54.80	33.71
SAT	**55.62**	**70.95**	**61.55**	**41.46**

Table 8: **BOBSL dataset:** We demonstrate strong performance of the SAT model on this test set.

4.6. Qualitative analysis

Fig. 4 illustrates several test examples on BSL-1K$_{aligned}$. The timeline shows the ground truth alignment (S_{gt}), our prediction (S_{pred}), as well as the S^{+}_{audio} baseline, alongside a sample of video frames and the query subtitle text. While the shifted baseline S^{+}_{audio} provides an approximate position, it is largely unaligned. Our model effectively learns to attend to both visual and textual cues. A typical failure mode happens when the prior position encoding is significantly far from the ground truth (see Fig. 4 right). For additional qualitative examples on BSL Corpus and BOBSL, we refer to App. Sec. C.

5. Conclusion

We presented a Transformer-based approach to synchronise subtitles with sign language video content in interpreted data. We showed that knowledge of subtitle content is essential to effectively align subtitles to signing. We hope that our work will be a stepping stone to obtain video-subtitle pairs that allow training of unconstrained machine translation systems for sign languages. Furthermore, our approach is potentially applicable to other domains, such as temporal grounding of sentences. We refer to App. Sec. E for a discussion on the broader impact on the community.

Acknowledgements. This work was supported by EPSRC grant ExTol and a Royal Society Research Professorship. We thank Tom Monnier, Himel Chowdhury, Abhishek Dutta, Ashish Thandavan, Annelies Braffort, Michèle Gouiffès and Igor Garbuz for their help.

References

[1] BBC-Oxford British Sign Language Dataset https://www.robots.ox.ac.uk/~vgg/data/bobsl/, 2021. 7

[2] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Andrew Brown, Chuhan Zhang, Ernesto Coto, Necati Cihan Camgöz, Ben Saunders, Abhishek Dutta, Neil Fox, Richard Bowden, Bencie Woll, and Andrew Zisserman. Signer diarisation in the wild. *Technical Report*, 2021. 2, 6

[3] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *Proc. ECCV*, 2020. 2, 4, 5, 6

[4] P. Bojanowski, Rémi Lajugie, E. Grave, Francis R. Bach, I. Laptev, J. Ponce, and C. Schmid. Weakly-supervised alignment of video with text. In *ICCV*, 2015. 3

[5] M. Borg and K. P. Camilleri. Sign language detection "in the wild" with recurrent neural networks. *ICASSP*, 2019. 2

[6] Danielle Bragg, Oscar Koller, et al. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *ACM SIGACCESS*, 2019. 2

[7] Patrick Buehler, Mark Everingham, and Andrew Zisserman. Learning sign language by watching TV (using weakly aligned subtitles). In *Proc. CVPR*, 2009. 2

[8] Patrick Buehler, Mark Everingham, and Andrew Zisserman. Employing signed TV broadcasts for automated learning of British sign language. In *Workshop on the Representation and Processing of Sign Languages*, 2010. 2

[9] Hannah Bull, Michèle Gouiffès, and Annelies Braffort. Automatic segmentation of sign language into subtitle-units. In *ECCVW, Sign Language Recognition, Translation and Production (SLRTP)*, 2020. 2, 6, 8

[10] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. SubUNets: End-to-end hand shape and continuous sign language recognition. In *ICCV*, 2017. 3

[11] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *CVPR*, 2018. 2, 3

[12] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *CVPR*, 2020. 2, 3

[13] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 4, 7

[14] J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, 2017. 4

[15] Jingyuan Chen, Xinpeng Chen, L. Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *EMNLP*, 2018. 3

[16] Ka Leong Cheng, Zhaoyang Yang, Qifeng Chen, and Yu-Wing Tai. Fully convolutional networks for continuous sign language recognition. In *ECCV*, 2020. 3

[17] N. Cherniavsky, R. E. Ladner, and E. A. Riskin. Activity detection in conversational sign language video for mobile telecommunication. In *IEEE International Conference on Automatic Face Gesture Recognition*, 2008. 2

[18] Helen Cooper and Richard Bowden. Learning signs from subtitles: A weakly supervised approach to sign language recognition. In *CVPR*, 2009. 2

[19] Karan Desai and Justin Johnson. VirTex: Learning visual representations from textual annotations. *arXiv:2006.06666*, 2021. 4

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2019. 4

[21] M. Everingham, Josef Sivic, and Andrew Zisserman. Hello! my name is... buffy" – automatic naming of characters in tv video. In *BMVC*, 2006. 3

[22] Iva Farag and Heike Brock. Learning motion disfluencies for automatic sign language segmentation. In *ICASSP*, 2019. 2

[23] Ali Farhadi and David Forsyth. Aligning ASL for statistical translation using a discriminative word model. In *CVPR*, 2006. 2, 6

[24] J. Fenlon. *Seeing sentence boundaries: the production and perception of visual markers signalling boundaries in signed languages*. PhD thesis, UCL, 2010. 2

[25] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, 2017. 3

[26] Binyam Gebrekidan Gebre, Peter Wittenburg, Tom Heskes, and Sebastian Drude. Motion history images for online speaker/signer diarization. In *ICASSP*, 2014. 2

[27] Binyam Gebrekidan Gebre, Peter Wittenburg, and Tom Heskes. Automatic signer diarization-the mover is the signer approach. In *CVPRW*, 2013. 2

[28] S. Ghosh, A. Agarwal, Zarana Parekh, and A. Hauptmann. Excl: Extractive clip localization using natural language descriptions. In *NAACL-HLT*, 2019. 3

[29] D. He, Xiang Zhao, Jizhou Huang, F. Li, Xiao Liu, and Shilei Wen. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *AAAI*, 2019. 3

[30] Lisa Anne Hendricks, O. Wang, E. Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. Localizing moments in video with natural language. In *ICCV*, 2017. 3

[31] Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. Video-based sign language recognition without temporal segmentation. In *AAAI*, 2018. 3

[32] Oscar Koller. Quantitative survey of the state of the art in sign language recognition. *arXiv:2008.09918*, 2020. 2

[33] Oscar Koller, O Zargaran, Hermann Ney, and Richard Bowden. Deep sign: Hybrid CNN-HMM for continuous sign language recognition. In *BMVC*, 2016. 3

[34] Oscar Koller, Sepehr Zargaran, and Hermann Ney. Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs. In *CVPR*, 2017. 3

[35] Dongxu Li, Chenchen Xu, Xin Yu, K. Zhang, Ben Swift, Hanna Suominen, and H. Li. TSPNet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. *NeurIPS*, 2020. 3

[36] Dongxu Li, Xin Yu, Chenchen Xu, Lars Petersson, and Hongdong Li. Transferring cross-domain knowledge for video sign language recognition. In *CVPR*, 2020. 2

[37] Rung-Huei Liang and Ming Ouhyoung. A real-time continuous gesture recognition system for sign language. In *Proceedings third IEEE international conference on automatic face and gesture recognition*, 1998. 2

[38] M. Liu, Xiang Wang, L. Nie, X. He, B. Chen, and Tat-Seng Chua. Attentive moment retrieval in videos. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018. 3

[39] Liliane Momeni, Gül Varol, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. Watch, read and lookup: Learning to spot signs from multiple supervisors. In *Proc. ACCV*, 2020. 2, 4, 5, 6

[40] Amit Moryossef, Ioannis Tsochantaridis, Roee Aharoni, Sarah Ebling, and Srini Narayanan. Real-Time Sign Language Detection using Human Pose Estimation. In *ECCVW, Sign Language Recognition, Translation and Production (SLRTP)*, 2020. 2

[41] C. Myers and L. Rabiner. A comparative study of several dynamic time-warping algorithms for connected-word recognition. *The Bell System Technical Journal*, 60:1389–1409, 1981. 3, 4

[42] Eng-Jon Ong, Oscar Koller, Nicolas Pugeault, and Richard Bowden. Sign spotting using hierarchical sequential patterns with temporal intervals. In *CVPR*, 2014. 2

[43] Tomas Pfister, James Charles, and Andrew Zisserman. Large-scale learning of sign language by watching TV (using co-occurrences). In *Proc. BMVC*, 2013. 2

[44] Katrin Renz, Nicolaj Stache, Samuel Albanie, and Gül Varol. Sign segmentation with temporal convolutional networks. In *International Conference on Acoustics, Speech, and Signal Processing*, 2021. 2

[45] K. P. Sankar, C. Jawahar, and Andrew Zisserman. Subtitle-free movie to script alignment. In *BMVC*, 2009. 3

[46] P. Santemiz, Oya Aran, M. Saraçlar, and L. Akarun. Automatic sign segmentation from continuous signing via multiple sequence alignment. *ICCVW*, 2009. 2

[47] Adam Schembri, Jordan Fenlon, Ramas Rentelis, and Kearsy Cormier. British Sign Language Corpus Project: A corpus of digital video data and annotations of British Sign Language 2008-2017 (Third Edition), 2017. 5, 7

[48] Adam Schembri, Jordan Fenlon, Ramas Rentelis, Sally Reynolds, and Kearsy Cormier. Building the British Sign Language Corpus. *Language Documentation & Conservation*, 7:136–154, 2013. 5, 7

[49] F. Shipman, Satyakiran Duggina, Caio D. D. Monteiro, and R. Gutierrez-Osuna. Speed-accuracy tradeoffs for detecting sign language content in video sharing sites. *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, 2017. 2

[50] Rachel Sutton-Spence and Bencie Woll. *The Linguistics of British Sign Language: An Introduction*. Cambridge University Press, 1999. 1, 2

[51] Makarand Tapaswi, M. Bäuml, and R. Stiefelhagen. Story-based video retrieval in tv series using plot synopses. In *Proceedings of International Conference on Multimedia Retrieval*, 2014. 3

[52] Makarand Tapaswi, M. Bäuml, and R. Stiefelhagen. Book2Movie: Aligning video scenes with book chapters. In *CVPR*, 2015. 3

[53] Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. Read and attend: Temporal localisation in sign language videos. In *CVPR*, 2021. 2, 4, 5

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4

[55] Weining Wang, Yan Huang, and Liang Wang. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *CVPR*, 2019. 3

[56] Huijuan Xu, Kun He, Bryan A. Plummer, L. Sigal, S. Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *AAAI*, 2019. 3

[57] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018. 6

[58] Hee-Deok Yang, Stan Sclaroff, and Seong-Whan Lee. Sign language spotting with a threshold model based on conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008. 2

[59] Ruiduo Yang and Sudeep Sarkar. Detecting coarticulation in sign language using conditional random fields. In *ICPR*, 2006. 2

[60] Yitian Yuan, T. Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *AAAI*, 2019. 3

[61] Runhao Zeng, H. Xu, W. Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *CVPR*, 2020. 3